

MCFD2 NUCLEIC ACIDS AND PROTEINS

This applications claims priority to co-pending U.S. provisional application serial
5 no. 60/448,264, filed 02/18/03, which is incorporated herein by reference in its entirety.
The present invention was made, in part, under research funded by Grant No. PO1 HL
57346. The government may have certain rights in the invention.

FIELD OF THE INVENTION

10 The present invention relates to early secretory pathway molecules, in particular
to the MCFD2 (multiple coagulation factor deficiency 2) protein, and nucleic acids
encoding the MCFD2 protein. The present invention provides assays for the detection of
MCFD2 and for *MCFD2* polymorphisms and mutations associated with disease states
and provides screening assays for the identification and use of compounds that alter
15 MCFD2 activity and/or biological pathways involving MCFD2.

BACKGROUND OF THE INVENTION

Combined deficiency of factor V and factor VIII (F5F8D, OMIM accession
number 227300) is associated with a bleeding tendency and plasma levels of FV and
20 FVIII in the range of 5%-30% of normal. Genetic deficiency of FVIII results in classic
hemophilia (hemophilia A) whereas inherited FV deficiency leads to parahemophilia, a
rare autosomal recessive condition exhibiting a similar hemorrhagic phenotype
(Ginsburg, 2002). Inheritance of F5F8D is autosomal recessive, representing a condition
distinct from the coinheritance of both FV deficiency and FVIII deficiency.

25 Mutations in *LMAN1*, which encodes the endoplasmic reticulum-Gogli
intermediate compartment type I membrane protein ERGIC-53, account for the majority
of patients with F5F8D disease (Nichols et al., 1998). However, no *LMAN1* mutations
can be identified in ~30% of affected families (Neerman-Arbez et al., 1999; Nichols et
al., 1999). For at least two of these families, F5F8D is not linked to the *LMAN1* locus
30 (Nichols et al., 1999). In addition, affected individuals from another two families with no
identified *LMAN1* mutation exhibit normal levels of *LMAN1* protein expression

(Neerman-Arbez et al., 1999). Taken together, these results indicate the existence of at least one additional locus responsible for F5F8D. The identical phenotypes observed in all F5F8D patients indicate that this second locus encodes a key component of the FV/FVIII secretory pathway.

5 The only treatment for F5F8D is factor replacement therapy. There is no known cure. Clearly there is a great need for identification of the molecular basis of F5F8D, as well as for improved diagnostics and treatments for F5F8D.

SUMMARY OF THE INVENTION

10 The present invention relates to early secretory pathway molecules, in particular to the MCFD2 (multiple coagulation factor deficiency 2) protein, and nucleic acids encoding the MCFD2 protein. The present invention provides assays for the detection of MCFD2 and for *MCFD2* polymorphisms and mutations associated with disease states and provides screening assays for the identification and use of compounds that alter
15 MCFD2 activity and/or biological pathways involving MCFD2.

 Accordingly, in some embodiments, the present invention provides an isolated and purified nucleic acid comprising a sequence encoding a protein of an MCFD2 wild type or mutant sequence described herein (e.g., SEQ ID NOs: 2, 6, 8, 10, 12, 14, 16, and 18). In some embodiments, the sequence is operably linked to a heterologous promoter.
20 In some embodiments, the sequence is contained within a vector. In some embodiments, the vector is within a host cell. In some embodiments, the present invention provides a computer readable medium encoding a representation of the nucleic acid sequence.

 The present invention also provides an isolated and purified nucleic acid sequence that hybridizes under conditions of low stringency to a MCFD2 nucleic acid described
25 herein (e.g., SEQ ID NOs: 1, 5, 7, 9, 11, 13, 15, and 17). In some embodiments, the nucleic acid sequence, alone or in combination with other reagents, hybridizes or otherwise detects a mutant MCFD2 sequence but not a wild-type sequence. In some embodiments, the sequence is contained within a vector. In some embodiments, the vector is in a host cell. In some embodiments, the host cell is located in an organism,
30 wherein the organism is a non-human animal.

The present invention additionally provides a protein encoded by a nucleic acid of SEQ ID NOs:1 and variants thereof that are at least 80% identical to SEQ ID NOs: 1, 5, 7, 9, 11, 13, 15, and 17. In some embodiments, the protein encoded by a nucleic acid that is at least 90%, and preferably at least 95% identical to SEQ ID NOs: 1, 5, 7, 9, 11, 13,
5 15, and 17.

The present invention further provides a composition comprising a nucleic acid that inhibits the binding of at least a portion of a nucleic acid selected from the group consisting of SEQ ID NOs:1, 5, 7, 9, 11, 13, 15, and 17 to their complementary sequences. In other embodiments, the present invention provides a polynucleotide
10 sequence comprising at least fifteen nucleotides capable of hybridizing under stringent conditions to the isolated nucleotide sequence.

In yet other embodiments, the present invention provides a composition comprising a variant MCFD2 polypeptide, wherein the polypeptide comprises a C-terminal truncation of SEQ ID NO:2. In some embodiments, the presence of the variant
15 polypeptide in a subject is indicative of Factor V and Factor VIII deficiency (F5F8D) disease in the subject.

In still further embodiments, the present invention provides a method for detection of a variant MCFD2 polypeptide in a subject, comprising: providing a biological sample from a subject, wherein the biological sample comprises MCFD2
20 polypeptide; and detecting the presence or absence of a variant MCFD2 polypeptide in the biological sample. In some embodiments, the variant MCFD2 polypeptide is a C-terminal truncation of SEQ ID NO:2. In some embodiments, the presence of the variant MCFD2 polypeptide is indicative of F5F8D disease in the subject. In some
25 embodiments, the biological sample is selected from the group consisting of a blood sample, a tissue sample, a urine sample, and an amniotic fluid sample. In some embodiments, the subject is selected from the group consisting of an embryo, a fetus, a newborn animal, and a young animal. In some embodiments, the animal is a human. In some embodiments, the detection comprises differential antibody binding. In other
30 embodiments, the detection comprises a gel-free truncation test. In still other embodiments, the detection comprises a Western blot.

The present invention further provides a kit comprising a reagent for detecting the presence or absence of a variant MCFD2 polypeptide in a biological sample. In some embodiments, the kit further comprises instruction for using the kit for detecting the presence or absence of a variant MCFD2 polypeptide in a biological sample. In some
5 embodiments, the instructions comprise instructions required by the U.S. Food and Drug Agency for *in vitro* diagnostic kits. In some embodiments, the kit further comprises instructions for diagnosing F5F8D in the subject based on the presence or absence of the variant MCFD2 polypeptide. In some embodiments, the F5F8D disease is MCFD2 mediated. In some embodiments, the reagent is one or more antibodies. In some
10 embodiments, the antibodies comprise a first antibody that specifically binds to the C-terminus of the MCFD2 polypeptide and a second antibody that specifically binds to the N-terminus of the MCFD2 polypeptide. In other embodiments, the reagents comprise reagents for performing a gel-free truncation test. In some embodiments, the variant MCFD2 polypeptide is a C-terminal truncation of SEQ ID NO:2. In some embodiments,
15 the biological sample is selected from the group consisting of a blood sample, a tissue sample, a urine sample, and an amniotic fluid sample.

The present invention also provides a method for screening compounds, comprising, providing a sample comprising MCFD2 and a test compound; and exposing the sample to the test compound and detecting a biological effect. In some embodiments,
20 the sample comprises a cell or a tissue (e.g., *in vitro*, *ex vivo*, or *in vivo*). In some embodiments, the biological effect is a change in activity of MCFD2 (measured directly or indirectly), a change in expression of MCFD2 (measured directly or indirectly), or a change in a disease phenotype.

The present invention further provides a method for expressing MCFD2 (e.g.,
25 wild-type MCFD2) *in vivo*, comprising, providing a subject and an expression vector encoding MCFD2; and introducing the expression vector into the subject under conditions such that MCFD2 is expressed. In some embodiments, the subject comprises a non-human animal. In other embodiments, the expression vector comprises an adenoviral vector. In some embodiments, the method further comprises the step of
30 measuring MCFD2 expression levels.

DESCRIPTION OF THE FIGURES

Figure 1 shows pedigrees of the F5F8D families used for linkage analysis. Circles indicate females and squares indicate males. Affected individuals are indicated by solid symbols. Individuals III:1 and III:4 of family 6 are siblings. Forty-five markers (22 from the Genethon and Marshfield comprehensive genetic maps and 23 designed from sequence-repeat information available at the Human Genome Project working draft) were used for haplotype analysis, with nine markers shown here. Common chromosomes in consanguineous families 1-6 and chromosomes shared by affected siblings in families 7-10 are indicated in yellow. Haplotypes are arbitrarily assigned in families 7-10 since no parental DNA was available. Inferred genotypes are indicated in parentheses. Marker genotypes not determined are indicated as dashes. Recombination events in individuals 1IV-4, 7II-2 and 7II-1 place the responsible gene below marker BZ30, BZ31, and BZ31, respectively, while a recombination event in individual 9II-1 places the gene above marker BZ18.

Figure 2 shows identification of the *MCFD2* gene. Figure 2A, the genetic and physical map of chromosome 2 in the interval between markers D2S2174 and D2S337. The ~1.8-cM nonrecombinant interval is shown in red. Major sequence gaps in the public genome assembly are denoted by blue rectangles. Transcripts localized to this interval are depicted by alternating black and grey bars. The *MCFD2* gene is indicated with an asterisk. The reference bar represents 1 cM. Figure 2B, the intron–exon structure of *MCFD2*. The coding region is illustrated in grey and 5' and 3'untranslated regions in green. Intron sizes are not drawn to scale. Splice mutations are depicted underneath the corresponding exons (blue diamonds). Scale bar, 200 nucleotides. Figure 2C, the predicted domain structure of MCFD2. The predicted signal peptide (predicted using SignalP version 1.1 (at the cbs.dtu.dk/services/SignalP web site) is indicated in yellow, and the two EF-hand domains in dark blue. The green triangles indicate the locations of the 2 missense mutations and red squares indicate frameshift mutations. Scale bar, 5 amino acids.

Figure 3 shows expression of MCFD2 in EBV-immortalized lymphocytes established from patients with F5F8D. Cell extracts (from 2x10⁵ cells) prepared from EBV-immortalized lymphoblasts from two normal control individuals (WT), two affected

individuals with *LMAN1* null mutations (*LMAN1*⁻, derived from family A19 and A12 (Neerman-Arbez et al., 1999)), four individuals with *MCFD2* mutations (*MCFD2*⁻, from left to right: family 11, family 7 II-1, family 12, family 6 IV-4), were separated by 15% SDS-PAGE and transferred to nitrocellulose membranes. Immunoblots were detected
5 with a monoclonal antibody against LMAN1 (Figure 2A) or a monoclonal antibody against MCFD2 (Figure 2B). Numbers to the left indicate the molecular weight (kDa) of pre-stained markers. Figure 2C, cell extracts (from 5x10⁶ cells) were immunoprecipitated with Sepharose beads conjugated to a rabbit anti-MCFD2 antisera and detected on a Western blot by a monoclonal antibody against MCFD2.

10 Figure 4 shows Northern and RACE Analysis of *MCFD2*. Figure 4A, Northern blot of polyA mRNA (2 mg/lane) from multiple human tissues was hybridized with a probe spanning the entire coding region of *MCFD2*. A 4.1-kb mRNA can be seen in all the tissues, although faint in brain and lung. Smaller transcripts in the range of 0.8~1.8-kb are also detected in some tissues. Figure 4B, the 5' untranslated region of the *MCFD2*
15 gene. Arrows indicate transcriptional start sites as determined by 5' RACE.

Figure 5 shows the intracellular localization of MCFD2. HeLa cells were transfected with pcDNA3.1-*myc-his* vectors expressing wild-type (WT), the D129E and I136T mutants, or the I136V substitution. Cells were stained at 24 hours post-transfection with a rabbit anti-*myc* antibody and either a monoclonal anti-LMAN1, or a
20 monoclonal anti-PDI antibody. Protein localization was detected by immunofluorescence confocal microscopy using a FITC-conjugated anti-mouse IgG and a Texas red-conjugated anti-rabbit IgG. Each group of four panels shows from the left, the green fluorescence from anti-LMAN1, the red fluorescence from anti-*myc*, the merged image from the two, and the merged image from anti-PDI (green) and anti-*myc*
25 (red).

Figure 6 shows that MCFD2 interacts with LMAN1. Figure 6A, COS-1 cells were either mock-transfected, or transfected with plasmids expressing wild-type MCFD2-*myc*, MCFD2 (D129E)-*myc*, MCFD2 (I136V)-*myc*, MCFD2 (I136T)-*myc*, and metabolically labeled with [35S]-methionine. Cell extracts were immunoprecipitated (IP)
30 with an anti-*myc* antibody (M) or an anti-LMAN1 antibody (L). Arrows (from top to bottom) denote LMAN1, MCFD2-*myc*, and endogenous MCFD2. Figure 6B, COS-1

cells were co-transfected with a plasmid expressing wild-type MCFD2-*myc* and an LMAN1-expression plasmid. Cell extracts were immunoprecipitated (IP) with an anti-*myc* antibody (M) or an anti-LMAN1 antibody (L). Where indicated (+), EGTA (2.5 mM) was added to the cell extract before the addition of antibodies. Calcium chloride (Ca, 10 mM) was added back to half of the EGTA-treated lysate, cleared by centrifugation, before the addition of antibodies. Arrows (from top to bottom) denote LMAN1 and MCFD2-*myc*.

Figure 7 shows the full length cDNA/mRNA of wild-type MCFD2 (SEQ ID NO:1).

Figure 8 shows the full length amino acid sequence of wild-type MCFD2 (SEQ ID NO:2)

Figure 9 shows the wild-type MCFD2 nucleic acid and amino acid sequences (SEQ ID NOs:1 and 2).

Figure 10 shows the genomic sequence encoding MCFD2 with exons underlined (SEQ ID NO:19).

GENERAL DESCRIPTION OF THE INVENTION

The present invention provides methods and compositions for detecting, characterizing, and/or modulating MCFD2 nucleic acids and proteins. In some embodiments, polymorphisms of MCFD2 are detected (directly or indirectly) and associated with disease states. In other embodiments, MCFD2 activity is altered in cells in vitro or in vivo to characterize MCFD2 expression or biological activity, to alter MCFD2 expression or biological activity (e.g., to characterize or treat F5F8D), and/or to identify compounds that alter MCFD2 expression or biological activity.

A cargo-specific ER-to-Golgi transport pathway has been identified by detection of mutations in *ERGIC-53 (LMAN1)* as a cause or combined deficiency of factor V and factor VIII (F5F8D) in several families (Nichols et al., 1998). Coagulation factor V (FV) and factor VIII (FVIII) are both required for the efficient function of the blood-clotting system. Combined deficiency of factor V and factor VIII (F5F8D, OMIM accession number 227300) is associated with a bleeding tendency and plasma levels of FV and FVIII in the range of 5%-30% of normal.

Experiments conducted during the course of development of the present invention identified mutations in the gene *MCFD2* as a second cause of F5F8D. *MCFD2* encodes a novel ER resident protein that is retained within the ER through its interaction with LMAN1. Multiple homozygous inactivating mutations in 9 families clearly establish
5 *MCFD2* as a second gene responsible for F5F8D (Table 1.). Five of the seven identified mutations are each unique to a single family, with each of the remaining two splice mutations occurring in two apparently unrelated families. These data demonstrate a role for MCFD2 in ER to Golgi transport for a specific subset of glycoproteins including factors V and VIII. The colocalization of MCFD2 and LMAN1 to the ERGIC (Figure 5),
10 as well as the direct interaction demonstrated by co-immunoprecipitation (Figure 6) are consistent with joint function of these two proteins in a common secretory pathway. In some embodiments of the present invention, MCFD2 is detected or modulated in conjunction with one or more other factors (e.g., LMAN1) with, for example, a shared biological pathway (i.e., multiplex detection or modulation).

15 Most of the MCFD2 mutations identified here (Table 1) result in complete loss of MCFD2 protein expression, similar to the spectrum of LMAN1 mutations previously observed in F5F8D (Nichols et al., 1998; Neerman-Arbez et al., 1999; Nichols et al., 1999). No MCFD2 protein is detectable by Western blot analysis in four patient lymphoblast cell lines available for study (Figure 3). While the present invention is not
20 limited to any particular mechanism and an understanding of the mechanism is not necessary to practice the present invention, it is contemplated that, taken together, these data suggest that low levels of MCFD2 and LMAN1 function are sufficient for FV and FVIII secretion, accounting for bias toward null mutations in the F5F8D patients studied to date, all of whom have been identified on the basis of clinical bleeding.

25 MCFD2 is partially homologous to a protein of unknown function, identified because of the presence of a transcribed transposon element sequence in the 3' untranslated region (Deka et al., 1988) (*HUMTRANSC*, GenBank #M23161). Analysis of the >100 EST sequences corresponding to full-length CDF2 suggests at least seven distinct alternatively spliced forms, including several that encode alternative forms of the
30 protein. The major mRNA species seen on Northern blot is ~4.1 kb in length (Figure

3A), corresponding to the transposon-containing transcript. However, three smaller bands are seen in several other tissues, which may represent alternatively spliced mRNA.

MCFD2 homologues with >80% amino acid identity are evident in other mammalian species including mouse and rat. Only the human gene contains the THE1 transposable element. Probable homologues of MCFD2 are also evident in multiple other vertebrate and invertebrate species, including zebrafish (~65% amino acid identity), *D. melanogaster* (42%), and *C. elegans* (45%).

The limited subset of proteins clinically affected by MCFD2/LMAN1 deficiency make this pathway an attractive therapeutic target for anticoagulation, particularly in light of the prothrombotic risk associated with FV Leiden and with elevated levels of FVIII. The present invention is not limited to a particular mechanism of action. Indeed, an understanding of the mechanism is not necessary to practice the present invention. Nonetheless, it is likely that both MCFD2 and LMAN1, interact within a novel shared pathogenic pathway. Thus, the present invention provides a novel gene with a critical role in ER to Golgi transport.

DEFINITIONS

To facilitate understanding of the invention, a number of terms are defined below.

As used herein, the term "MCFD2" or "combined factor deficiency 2" when used in reference to a protein or nucleic acid refers to a protein or nucleic acid encoding a protein that, in some mutant forms, is correlated with factor V factor VIII deficiency (F5F8D). The term MCFD2 encompasses both proteins that are identical to wild-type MCFD2 and those that are derived from wild type MCFD2 (*e.g.*, variants of MCFD2 or chimeric genes constructed with portions of MCFD2 coding regions). In some embodiments, the "MCFD2" is the wild type nucleic acid (SEQ ID NO: 1) or amino acid (SEQ ID NO:2) sequence. In other embodiments, the "MCFD2" is a variant or mutant (*e.g.*, including, but not limited to, the nucleic acid sequences described by SEQ ID NOS: 5, 7, 9, 11, 13, 15, and 17 and the amino acid sequences described by SEQ ID NOS: 6, 8, 10, 12, 14, 16, and 18).

As used herein, the term "C-terminal truncation of SEQ ID NO:2" refers to a polypeptide comprising a portion of SEQ ID NO:2, wherein the portion comprises the N-

terminus of SEQ ID NO:2. In preferred embodiments, the N-terminal portion comprises at least 200 amino acids, preferably at least 400 amino acids, and even more preferably at least 700 amino acids of SEQ ID NO:2.

As used herein, the term "instructions for using said kit for said detecting the presence or absence of a variant MCFD2 polypeptide in a said biological sample" includes instructions for using the reagents contained in the kit for the detection of variant and wild type MCFD2 polypeptides. In some embodiments, the instructions further comprise the statement of intended use required by the U.S. Food and Drug Administration (FDA) in labeling *in vitro* diagnostic products. The FDA classifies *in vitro* diagnostics as medical devices and requires that they be approved through the 510(k) procedure. Information required in an application under 510(k) includes: 1) The *in vitro* diagnostic product name, including the trade or proprietary name, the common or usual name, and the classification name of the device; 2) The intended use of the product; 3) The establishment registration number, if applicable, of the owner or operator submitting the 510(k) submission; the class in which the *in vitro* diagnostic product was placed under section 513 of the FD&C Act, if known, its appropriate panel, or, if the owner or operator determines that the device has not been classified under such section, a statement of that determination and the basis for the determination that the *in vitro* diagnostic product is not so classified; 4) Proposed labels, labeling and advertisements sufficient to describe the *in vitro* diagnostic product, its intended use, and directions for use. Where applicable, photographs or engineering drawings should be supplied; 5) A statement indicating that the device is similar to and/or different from other *in vitro* diagnostic products of comparable type in commercial distribution in the U.S., accompanied by data to support the statement; 6) A 510(k) summary of the safety and effectiveness data upon which the substantial equivalence determination is based; or a statement that the 510(k) safety and effectiveness information supporting the FDA finding of substantial equivalence will be made available to any person within 30 days of a written request; 7) A statement that the submitter believes, to the best of their knowledge, that all data and information submitted in the premarket notification are truthful and accurate and that no material fact has been omitted; 8) Any additional information regarding the *in vitro* diagnostic product requested that is necessary for the

FDA to make a substantial equivalency determination. Additional information is available at the Internet web page of the U.S. FDA.

The term "gene" refers to a nucleic acid (*e.g.*, DNA) sequence that comprises coding sequences necessary for the production of a polypeptide, RNA (*e.g.*, including but not limited to, mRNA, tRNA and rRNA) or precursor (*e.g.*, MCFD2). The polypeptide, RNA, or precursor can be encoded by a full length coding sequence or by any portion of the coding sequence so long as the desired activity or functional properties (*e.g.*, enzymatic activity, ligand binding, signal transduction, etc.) of the full-length or fragment are retained. The term also encompasses the coding region of a structural gene and the including sequences located adjacent to the coding region on both the 5' and 3' ends for a distance of about 1 kb on either end such that the gene corresponds to the length of the full-length mRNA. The sequences that are located 5' of the coding region and which are present on the mRNA are referred to as 5' untranslated sequences. The sequences that are located 3' or downstream of the coding region and that are present on the mRNA are referred to as 3' untranslated sequences. The term "gene" encompasses both cDNA and genomic forms of a gene. A genomic form or clone of a gene contains the coding region interrupted with non-coding sequences termed "introns" or "intervening regions" or "intervening sequences." Introns are segments of a gene that are transcribed into nuclear RNA (hnRNA); introns may contain regulatory elements such as enhancers. Introns are removed or "spliced out" from the nuclear or primary transcript; introns therefore are absent in the messenger RNA (mRNA) transcript. The mRNA functions during translation to specify the sequence or order of amino acids in a nascent polypeptide.

In particular, the term "MCFD2 gene" refers to the full-length MCFD2 nucleotide sequence (*e.g.*, contained in SEQ ID NO: 1). However, it is also intended that the term encompass fragments of the MCFD2 sequence, mutants as well as other domains within the full-length MCFD2 nucleotide sequence. Furthermore, the terms "MCFD2 nucleotide sequence" or "MCFD2 polynucleotide sequence" encompasses DNA, cDNA, and RNA (*e.g.*, mRNA) sequences.

Where "amino acid sequence" is recited herein to refer to an amino acid sequence of a naturally occurring protein molecule, "amino acid sequence" and like terms, such as

"polypeptide" or "protein" are not meant to limit the amino acid sequence to the complete, native amino acid sequence associated with the recited protein molecule.

In addition to containing introns, genomic forms of a gene may also include sequences located on both the 5' and 3' end of the sequences that are present on the RNA transcript. These sequences are referred to as "flanking" sequences or regions (these flanking sequences are located 5' or 3' to the non-translated sequences present on the mRNA transcript). The 5' flanking region may contain regulatory sequences such as promoters and enhancers that control or influence the transcription of the gene. The 3' flanking region may contain sequences that direct the termination of transcription, post-transcriptional cleavage and polyadenylation.

The term "wild-type" refers to a gene or gene product that has the characteristics of that gene or gene product when isolated from a naturally occurring source. A wild-type gene is that which is most frequently observed in a population and is thus arbitrarily designated the "normal" or "wild-type" form of the gene. In contrast, the terms "modified," "mutant," "polymorphism," and "variant" refer to a gene or gene product that displays modifications in sequence and/or functional properties (*i.e.*, altered characteristics) when compared to the wild-type gene or gene product. It is noted that naturally-occurring mutants can be isolated; these are identified by the fact that they have altered characteristics when compared to the wild-type gene or gene product.

As used herein, the terms "nucleic acid molecule encoding," "DNA sequence encoding," and "DNA encoding" refer to the order or sequence of deoxyribonucleotides along a strand of deoxyribonucleic acid. The order of these deoxyribonucleotides determines the order of amino acids along the polypeptide (protein) chain. The DNA sequence thus codes for the amino acid sequence.

DNA molecules are said to have "5' ends" and "3' ends" because mononucleotides are reacted to make oligonucleotides or polynucleotides in a manner such that the 5' phosphate of one mononucleotide pentose ring is attached to the 3' oxygen of its neighbor in one direction via a phosphodiester linkage. Therefore, an end of an oligonucleotides or polynucleotide, referred to as the "5' end" if its 5' phosphate is not linked to the 3' oxygen of a mononucleotide pentose ring and as the "3' end" if its 3' oxygen is not linked to a 5' phosphate of a subsequent mononucleotide pentose ring. As used herein, a nucleic acid

sequence, even if internal to a larger oligonucleotide or polynucleotide, also may be said to have 5' and 3' ends. In either a linear or circular DNA molecule, discrete elements are referred to as being "upstream" or 5' of the "downstream" or 3' elements. This terminology reflects the fact that transcription proceeds in a 5' to 3' fashion along the DNA strand. The promoter and enhancer elements that direct transcription of a linked gene are generally located 5' or upstream of the coding region. However, enhancer elements can exert their effect even when located 3' of the promoter element and the coding region. Transcription termination and polyadenylation signals are located 3' or downstream of the coding region.

As used herein, the terms "an oligonucleotide having a nucleotide sequence encoding a gene" and "polynucleotide having a nucleotide sequence encoding a gene," means a nucleic acid sequence comprising the coding region of a gene or, in other words, the nucleic acid sequence that encodes a gene product. The coding region may be present in a cDNA, genomic DNA, or RNA form. When present in a DNA form, the oligonucleotide or polynucleotide may be single-stranded (*i.e.*, the sense strand) or double-stranded. Suitable control elements such as enhancers/promoters, splice junctions, polyadenylation signals, *etc.* may be placed in close proximity to the coding region of the gene if needed to permit proper initiation of transcription and/or correct processing of the primary RNA transcript. Alternatively, the coding region utilized in the expression vectors of the present invention may contain endogenous enhancers/promoters, splice junctions, intervening sequences, polyadenylation signals, *etc.* or a combination of both endogenous and exogenous control elements.

As used herein, the term "regulatory element" refers to a genetic element that controls some aspect of the expression of nucleic acid sequences. For example, a promoter is a regulatory element that facilitates the initiation of transcription of an operably linked coding region. Other regulatory elements include splicing signals, polyadenylation signals, termination signals, *etc.*

As used herein, the terms "complementary" or "complementarity" are used in reference to polynucleotides (*i.e.*, a sequence of nucleotides) related by the base-pairing rules. For example, for the sequence 5'-"A-G-T-3'," is complementary to the sequence 3'-"T-C-A-5'." Complementarity may be "partial," in which only some of the nucleic acids'

bases are matched according to the base pairing rules. Or, there may be "complete" or "total" complementarity between the nucleic acids. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands. This is of particular importance in amplification reactions, as well as detection methods that depend upon binding between nucleic acids.

The term "homology" refers to a degree of complementarity. There may be partial homology or complete homology (*i.e.*, identity). A partially complementary sequence is one that at least partially inhibits a completely complementary sequence from hybridizing to a target nucleic acid and is referred to using the functional term "substantially homologous." The term "inhibition of binding," when used in reference to nucleic acid binding, refers to inhibition of binding caused by competition of homologous sequences for binding to a target sequence. The inhibition of hybridization of the completely complementary sequence to the target sequence may be examined using a hybridization assay (Southern or Northern blot, solution hybridization and the like) under conditions of low stringency. A substantially homologous sequence or probe will compete for and inhibit the binding (*i.e.*, the hybridization) of a completely homologous to a target under conditions of low stringency. This is not to say that conditions of low stringency are such that non-specific binding is permitted; low stringency conditions require that the binding of two sequences to one another be a specific (*i.e.*, selective) interaction. The absence of non-specific binding may be tested by the use of a second target that lacks even a partial degree of complementarity (*e.g.*, less than about 30% identity); in the absence of non-specific binding the probe will not hybridize to the second non-complementary target.

The art knows well that numerous equivalent conditions may be employed to comprise low stringency conditions; factors such as the length and nature (DNA, RNA, base composition) of the probe and nature of the target (DNA, RNA, base composition, present in solution or immobilized, etc.) and the concentration of the salts and other components (*e.g.*, the presence or absence of formamide, dextran sulfate, polyethylene glycol) are considered and the hybridization solution may be varied to generate conditions of low stringency hybridization different from, but equivalent to, the above

listed conditions. In addition, the art knows conditions that promote hybridization under conditions of high stringency (*e.g.*, increasing the temperature of the hybridization and/or wash steps, the use of formamide in the hybridization solution, etc.).

When used in reference to a double-stranded nucleic acid sequence such as a
5 cDNA or genomic clone, the term "substantially homologous" refers to any probe that can hybridize to either or both strands of the double-stranded nucleic acid sequence under conditions of low stringency as described above.

A gene may produce multiple RNA species that are generated by differential splicing of the primary RNA transcript. cDNAs that are splice variants of the same gene
10 will contain regions of sequence identity or complete homology (representing the presence of the same exon or portion of the same exon on both cDNAs) and regions of complete non-identity (for example, representing the presence of exon "A" on cDNA 1 wherein cDNA 2 contains exon "B" instead). Because the two cDNAs contain regions of sequence identity they will both hybridize to a probe derived from the entire gene or
15 portions of the gene containing sequences found on both cDNAs; the two splice variants are therefore substantially homologous to such a probe and to each other.

When used in reference to a single-stranded nucleic acid sequence, the term "substantially homologous" refers to any probe that can hybridize (*i.e.*, it is the complement of) the single-stranded nucleic acid sequence under conditions of low
20 stringency as described above.

As used herein, the term "competes for binding" is used in reference to a first polypeptide with an activity which binds to the same substrate as does a second polypeptide with an activity, where the second polypeptide is a variant of the first polypeptide or a related or dissimilar polypeptide. The efficiency (*e.g.*, kinetics or
25 thermodynamics) of binding by the first polypeptide may be the same as or greater than or less than the efficiency substrate binding by the second polypeptide. For example, the equilibrium binding constant (K_D) for binding to the substrate may be different for the two polypeptides. The term " K_m " as used herein refers to the Michaelis-Menton constant for an enzyme and is defined as the concentration of the specific substrate at which a
30 given enzyme yields one-half its maximum velocity in an enzyme catalyzed reaction.

As used herein, the term "hybridization" is used in reference to the pairing of complementary nucleic acids. Hybridization and the strength of hybridization (*i.e.*, the strength of the association between the nucleic acids) is impacted by such factors as the degree of complementarity between the nucleic acids, stringency of the conditions involved, the T_m of the formed hybrid, and the G:C ratio within the nucleic acids.

As used herein, the term " T_m " is used in reference to the "melting temperature."

The melting temperature is the temperature at which a population of double-stranded nucleic acid molecules becomes half dissociated into single strands. The equation for calculating the T_m of nucleic acids is well known in the art. As indicated by standard references, a simple estimate of the T_m value may be calculated by the equation: $T_m = 81.5 + 0.41(\% G + C)$, when a nucleic acid is in aqueous solution at 1 M NaCl (*See e.g.*, Anderson and Young, Quantitative Filter Hybridization, in *Nucleic Acid Hybridization* [1985]). Other references include more sophisticated computations that take structural as well as sequence characteristics into account for the calculation of T_m .

As used herein the term "stringency" is used in reference to the conditions of temperature, ionic strength, and the presence of other compounds such as organic solvents, under which nucleic acid hybridizations are conducted. Those skilled in the art will recognize that "stringency" conditions may be altered by varying the parameters just described either individually or in concert. With "high stringency" conditions, nucleic acid base pairing will occur only between nucleic acid fragments that have a high frequency of complementary base sequences (*e.g.*, hybridization under "high stringency" conditions may occur between homologs with about 85-100% identity, preferably about 70-100% identity). With medium stringency conditions, nucleic acid base pairing will occur between nucleic acids with an intermediate frequency of complementary base sequences (*e.g.*, hybridization under "medium stringency" conditions may occur between homologs with about 50-70% identity). Thus, conditions of "weak" or "low" stringency are often required with nucleic acids that are derived from organisms that are genetically diverse, as the frequency of complementary sequences is usually less.

"High stringency conditions" when used in reference to nucleic acid hybridization comprise conditions equivalent to binding or hybridization at 42 C in a solution

consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l NaH₂PO₄ H₂O and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.5% SDS, 5X Denhardt's reagent and 100 µg/ml denatured salmon sperm DNA followed by washing in a solution comprising 0.1X SSPE, 1.0% SDS at 42 C when a probe of about 500 nucleotides in length is employed.

5 "Medium stringency conditions" when used in reference to nucleic acid hybridization comprise conditions equivalent to binding or hybridization at 42 C in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l NaH₂PO₄ H₂O and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.5% SDS, 5X Denhardt's reagent and 100 µg/ml denatured salmon sperm DNA followed by washing in a solution comprising 1.0X SSPE,
10 1.0% SDS at 42 C when a probe of about 500 nucleotides in length is employed.

"Low stringency conditions" comprise conditions equivalent to binding or hybridization at 42 C in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l NaH₂PO₄ H₂O and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.1% SDS, 5X Denhardt's reagent [50X Denhardt's contains per 500 ml: 5 g Ficoll (Type 400,
15 Pharmacia), 5 g BSA (Fraction V; Sigma)] and 100 µg/ml denatured salmon sperm DNA followed by washing in a solution comprising 5X SSPE, 0.1% SDS at 42 C when a probe of about 500 nucleotides in length is employed. The present invention is not limited to the hybridization of probes of about 500 nucleotides in length. The present invention contemplates the use of probes between approximately 10 nucleotides up to several
20 thousand (*e.g.*, at least 5000) nucleotides in length.

One skilled in the relevant art understands that stringency conditions may be altered for probes of other sizes (See *e.g.*, Anderson and Young, Quantitative Filter Hybridization, in *Nucleic Acid Hybridization* [1985] and Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Press, NY [1989]).

25 The following terms are used to describe the sequence relationships between two or more polynucleotides: "reference sequence", "sequence identity", "percentage of sequence identity", and "substantial identity". A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA sequence
30 given in a sequence listing or may comprise a complete gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides

in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (*i.e.*, a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides, and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or
5 more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity. A "comparison window", as used herein, refers to a conceptual segment of at least 20 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 20 contiguous nucleotides and
10 wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (*i.e.*, gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith and
15 Waterman [Smith and Waterman, *Adv. Appl. Math.* 2: 482 (1981)] by the homology alignment algorithm of Needleman and Wunsch [Needleman and Wunsch, *J. Mol. Biol.* 48:443 (1970)], by the search for similarity method of Pearson and Lipman [Pearson and Lipman, *Proc. Natl. Acad. Sci. (U.S.A.)* 85:2444 (1988)], by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the
20 Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, Wis.), or by inspection, and the best alignment (*i.e.*, resulting in the highest percentage of homology over the comparison window) generated by the various methods is selected. The term "sequence identity" means that two polynucleotide sequences are identical (*i.e.*, on a nucleotide-by-nucleotide basis) over the window of
25 comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (*e.g.*, A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (*i.e.*, the window
30 size), and multiplying the result by 100 to yield the percentage of sequence identity. The terms "substantial identity" as used herein denotes a characteristic of a polynucleotide

sequence, wherein the polynucleotide comprises a sequence that has at least 85 percent sequence identity, preferably at least 90 to 95 percent sequence identity, more usually at least 99 percent sequence identity as compared to a reference sequence over a comparison window of at least 20 nucleotide positions, frequently over a window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison. The reference sequence may be a subset of a larger sequence, for example, as a segment of the full-length sequences of the compositions claimed in the present invention (*e.g.*, MCFD2).

As applied to polypeptides, the term "substantial identity" means that two peptide sequences, when optimally aligned, such as by the programs GAP or BESTFIT using default gap weights, share at least 80 percent sequence identity, preferably at least 90 percent sequence identity, more preferably at least 95 percent sequence identity or more (*e.g.*, 99 percent sequence identity). Preferably, residue positions that are not identical differ by conservative amino acid substitutions. Conservative amino acid substitutions refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are: valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

The term "fragment" as used herein refers to a polypeptide that has an amino-terminal and/or carboxy-terminal deletion as compared to the native protein, but where the remaining amino acid sequence is identical to the corresponding positions in the amino acid sequence deduced from a full-length cDNA sequence. Fragments typically are at least 4 amino acids long, preferably at least 20 amino acids long, usually at least 50

amino acids long or longer, and span the portion of the polypeptide required for intermolecular binding of the compositions (claimed in the present invention) with its various ligands and/or substrates.

5 The term "polymorphic locus" is a locus present in a population that shows variation between members of the population (*i.e.*, the most common allele has a frequency of less than 0.95). In contrast, a "monomorphic locus" is a genetic locus at little or no variations seen between members of the population (generally taken to be a locus at which the most common allele exceeds a frequency of 0.95 in the gene pool of the population).

10 As used herein, the term "genetic variation information" or "genetic variant information" refers to the presence or absence of one or more variant nucleic acid sequences (*e.g.*, polymorphism or mutations) in a given allele of a particular gene (*e.g.*, the MCFD2 gene).

15 As used herein, the term "detection assay" refers to an assay for detecting the presence of absence of variant nucleic acid sequences (*e.g.*, polymorphism or mutations) in a given allele of a particular gene (*e.g.*, the MCFD2 gene). Examples of suitable detection assays include, but are not limited to, those described below in Section III B.

20 The term "naturally-occurring" as used herein as applied to an object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally-occurring.

25 "Amplification" is a special case of nucleic acid replication involving template specificity. It is to be contrasted with non-specific template replication (*i.e.*, replication that is template-dependent but not dependent on a specific template). Template specificity is here distinguished from fidelity of replication (*i.e.*, synthesis of the proper polynucleotide sequence) and nucleotide (ribo- or deoxyribo-) specificity. Template specificity is frequently described in terms of "target" specificity. Target sequences are "targets" in the sense that they are sought to be sorted out from other nucleic acid.

30 Amplification techniques have been designed primarily for this sorting out.

Template specificity is achieved in most amplification techniques by the choice of enzyme. Amplification enzymes are enzymes that, under conditions they are used, will process only specific sequences of nucleic acid in a heterogeneous mixture of nucleic acid. For example, in the case of Q β replicase, MDV-1 RNA is the specific template for the replicase (D.L. Kacian *et al.*, Proc. Natl. Acad. Sci. USA 69:3038 [1972]). Other nucleic acid will not be replicated by this amplification enzyme. Similarly, in the case of T7 RNA polymerase, this amplification enzyme has a stringent specificity for its own promoters (Chamberlin *et al.*, Nature 228:227 [1970]). In the case of T4 DNA ligase, the enzyme will not ligate the two oligonucleotides or polynucleotides, where there is a mismatch between the oligonucleotide or polynucleotide substrate and the template at the ligation junction (D.Y. Wu and R. B. Wallace, Genomics 4:560 [1989]). Finally, *Taq* and *Pfu* polymerases, by virtue of their ability to function at high temperature, are found to display high specificity for the sequences bounded and thus defined by the primers; the high temperature results in thermodynamic conditions that favor primer hybridization with the target sequences and not hybridization with non-target sequences (H.A. Erlich (ed.), *PCR Technology*, Stockton Press [1989]).

As used herein, the term "amplifiable nucleic acid" is used in reference to nucleic acids that may be amplified by any amplification method. It is contemplated that "amplifiable nucleic acid" will usually comprise "sample template."

As used herein, the term "sample template" refers to nucleic acid originating from a sample that is analyzed for the presence of "target" (defined below). In contrast, "background template" is used in reference to nucleic acid other than sample template that may or may not be present in a sample. Background template is most often inadvertent. It may be the result of carryover, or it may be due to the presence of nucleic acid contaminants sought to be purified away from the sample. For example, nucleic acids from organisms other than those to be detected may be present as background in a test sample.

As used herein, the term "primer" refers to an oligonucleotide, whether occurring naturally as in a purified restriction digest or produced synthetically, which is capable of acting as a point of initiation of synthesis when placed under conditions in which synthesis of a primer extension product which is complementary to a nucleic acid strand

is induced, (*i.e.*, in the presence of nucleotides and an inducing agent such as DNA polymerase and at a suitable temperature and pH). The primer is preferably single stranded for maximum efficiency in amplification, but may alternatively be double stranded. If double stranded, the primer is first treated to separate its strands before being
5 used to prepare extension products. Preferably, the primer is an oligodeoxyribonucleotide. The primer must be sufficiently long to prime the synthesis of extension products in the presence of the inducing agent. The exact lengths of the primers will depend on many factors, including temperature, source of primer and the use of the method.

10 As used herein, the term "probe" refers to an oligonucleotide (*i.e.*, a sequence of nucleotides), whether occurring naturally as in a purified restriction digest or produced synthetically, recombinantly or by PCR amplification, that is capable of hybridizing to another oligonucleotide of interest. A probe may be single-stranded or double-stranded. Probes are useful in the detection, identification and isolation of particular gene
15 sequences. It is contemplated that any probe used in the present invention will be labeled with any "reporter molecule," so that is detectable in any detection system, including, but not limited to enzyme (*e.g.*, ELISA, as well as enzyme-based histochemical assays), fluorescent, radioactive, and luminescent systems. It is not intended that the present invention be limited to any particular detection system or label.

20 As used herein, the term "target," refers to a nucleic acid sequence or structure to be detected or characterized. Thus, the "target" is sought to be sorted out from other nucleic acid sequences. A "segment" is defined as a region of nucleic acid within the target sequence.

As used herein, the term "polymerase chain reaction" ("PCR") refers to the
25 method of K.B. Mullis U.S. Patent Nos. 4,683,195, 4,683,202, and 4,965,188, hereby incorporated by reference, that describe a method for increasing the concentration of a segment of a target sequence in a mixture of genomic DNA without cloning or purification. This process for amplifying the target sequence consists of introducing a large excess of two oligonucleotide primers to the DNA mixture containing the desired
30 target sequence, followed by a precise sequence of thermal cycling in the presence of a DNA polymerase. The two primers are complementary to their respective strands of the

double stranded target sequence. To effect amplification, the mixture is denatured and the primers then annealed to their complementary sequences within the target molecule. Following annealing, the primers are extended with a polymerase so as to form a new pair of complementary strands. The steps of denaturation, primer annealing, and
5 polymerase extension can be repeated many times (*i.e.*, denaturation, annealing and extension constitute one "cycle"; there can be numerous "cycles") to obtain a high concentration of an amplified segment of the desired target sequence. The length of the amplified segment of the desired target sequence is determined by the relative positions of the primers with respect to each other, and therefore, this length is a controllable
10 parameter. By virtue of the repeating aspect of the process, the method is referred to as the "polymerase chain reaction" (hereinafter "PCR"). Because the desired amplified segments of the target sequence become the predominant sequences (in terms of concentration) in the mixture, they are said to be "PCR amplified."

With PCR, it is possible to amplify a single copy of a specific target sequence in
15 genomic DNA to a level detectable by several different methodologies (*e.g.*, hybridization with a labeled probe; incorporation of biotinylated primers followed by avidin-enzyme conjugate detection; incorporation of ³²P-labeled deoxynucleotide triphosphates, such as dCTP or dATP, into the amplified segment). In addition to genomic DNA, any oligonucleotide or polynucleotide sequence can be amplified with the
20 appropriate set of primer molecules. In particular, the amplified segments created by the PCR process itself are, themselves, efficient templates for subsequent PCR amplifications.

As used herein, the terms "PCR product," "PCR fragment," and "amplification product" refer to the resultant mixture of compounds after two or more cycles of the PCR
25 steps of denaturation, annealing and extension are complete. These terms encompass the case where there has been amplification of one or more segments of one or more target sequences.

As used herein, the term "amplification reagents" refers to those reagents (deoxyribonucleotide triphosphates, buffer, etc.), needed for amplification except for
30 primers, nucleic acid template, and the amplification enzyme. Typically, amplification

reagents along with other reaction components are placed and contained in a reaction vessel (test tube, microwell, etc.).

As used herein, the terms "restriction endonucleases" and "restriction enzymes" refer to bacterial enzymes, each of which cut double-stranded DNA at or near a specific nucleotide sequence.

As used herein, the term "recombinant DNA molecule" as used herein refers to a DNA molecule that is comprised of segments of DNA joined together by means of molecular biological techniques.

As used herein, the term "antisense" is used in reference to RNA sequences that are complementary to a specific RNA sequence (*e.g.*, mRNA). Included within this definition are antisense RNA ("asRNA") molecules involved in gene regulation by bacteria. Antisense RNA may be produced by any method, including synthesis by splicing the gene(s) of interest in a reverse orientation to a viral promoter that permits the synthesis of a coding strand. Once introduced into an embryo, this transcribed strand combines with natural mRNA produced by the embryo to form duplexes. These duplexes then block either the further transcription of the mRNA or its translation. In this manner, mutant phenotypes may be generated. The term "antisense strand" is used in reference to a nucleic acid strand that is complementary to the "sense" strand. The designation (-) (*i.e.*, "negative") is sometimes used in reference to the antisense strand, with the designation (+) sometimes used in reference to the sense (*i.e.*, "positive") strand.

The term "isolated" when used in relation to a nucleic acid, as in "an isolated oligonucleotide" or "isolated polynucleotide" refers to a nucleic acid sequence that is identified and separated from at least one contaminant nucleic acid with which it is ordinarily associated in its natural source. Isolated nucleic acid is present in a form or setting that is different from that in which it is found in nature. In contrast, non-isolated nucleic acids are nucleic acids such as DNA and RNA found in the state they exist in nature. For example, a given DNA sequence (*e.g.*, a gene) is found on the host cell chromosome in proximity to neighboring genes; RNA sequences, such as a specific mRNA sequence encoding a specific protein, are found in the cell as a mixture with numerous other mRNAs that encode a multitude of proteins. However, isolated nucleic acid encoding MCFD2 includes, by way of example, such nucleic acid in cells ordinarily

expressing MCFD2 where the nucleic acid is in a chromosomal location different from that of natural cells, or is otherwise flanked by a different nucleic acid sequence than that found in nature. The isolated nucleic acid, oligonucleotide, or polynucleotide may be present in single-stranded or double-stranded form. When an isolated nucleic acid,
5 oligonucleotide or polynucleotide is to be utilized to express a protein, the oligonucleotide or polynucleotide will contain at a minimum the sense or coding strand (*i.e.*, the oligonucleotide or polynucleotide may single-stranded), but may contain both the sense and anti-sense strands (*i.e.*, the oligonucleotide or polynucleotide may be double-stranded).

10 As used herein, a "portion of a chromosome" refers to a discrete section of the chromosome. Chromosomes are divided into sites or sections by cytogeneticists as follows: the short (relative to the centromere) arm of a chromosome is termed the "p" arm; the long arm is termed the "q" arm. Each arm is then divided into 2 regions termed region 1 and region 2 (region 1 is closest to the centromere). Each region is further
15 divided into bands. The bands may be further divided into sub-bands. For example, the 11p15.5 portion of human chromosome 11 is the portion located on chromosome 11 (11) on the short arm (p) in the first region (1) in the 5th band (5) in sub-band 5 (.5). A portion of a chromosome may be "altered;" for instance the entire portion may be absent due to a deletion or may be rearranged (*e.g.*, inversions, translocations, expanded or
20 contracted due to changes in repeat regions). In the case of a deletion, an attempt to hybridize (*i.e.*, specifically bind) a probe homologous to a particular portion of a chromosome could result in a negative result (*i.e.*, the probe could not bind to the sample containing genetic material suspected of containing the missing portion of the chromosome). Thus, hybridization of a probe homologous to a particular portion of a
25 chromosome may be used to detect alterations in a portion of a chromosome.

The term "sequences associated with a chromosome" means preparations of chromosomes (*e.g.*, spreads of metaphase chromosomes), nucleic acid extracted from a sample containing chromosomal DNA (*e.g.*, preparations of genomic DNA); the RNA that is produced by transcription of genes located on a chromosome (*e.g.*, hnRNA and
30 mRNA), and cDNA copies of the RNA transcribed from the DNA located on a chromosome. Sequences associated with a chromosome may be detected by numerous

techniques including probing of Southern and Northern blots and *in situ* hybridization to RNA, DNA, or metaphase chromosomes with probes containing sequences homologous to the nucleic acids in the above listed preparations.

As used herein the term "portion" when in reference to a nucleotide sequence (as in "a portion of a given nucleotide sequence") refers to fragments of that sequence. The fragments may range in size from four nucleotides to the entire nucleotide sequence minus one nucleotide (10 nucleotides, 20, 30, 40, 50, 100, 200, etc.).

As used herein the term "coding region" when used in reference to structural gene refers to the nucleotide sequences that encode the amino acids found in the nascent polypeptide as a result of translation of a mRNA molecule. The coding region is bounded, in eukaryotes, on the 5' side by the nucleotide triplet "ATG" that encodes the initiator methionine and on the 3' side by one of the three triplets, which specify stop codons (*i.e.*, TAA, TAG, TGA).

As used herein, the term "purified" or "to purify" refers to the removal of contaminants from a sample. For example, MCFD2 antibodies are purified by removal of contaminating non-immunoglobulin proteins; they are also purified by the removal of immunoglobulin that does not bind MCFD2. The removal of non-immunoglobulin proteins and/or the removal of immunoglobulins that do not bind MCFD2 results in an increase in the percent of MCFD2-reactive immunoglobulins in the sample. In another example, recombinant MCFD2 polypeptides are expressed in bacterial host cells and the polypeptides are purified by the removal of host cell proteins; the percent of recombinant MCFD2 polypeptides is thereby increased in the sample.

The term "recombinant DNA molecule" as used herein refers to a DNA molecule that is comprised of segments of DNA joined together by means of molecular biological techniques.

The term "recombinant protein" or "recombinant polypeptide" as used herein refers to a protein molecule that is expressed from a recombinant DNA molecule.

The term "native protein" as used herein to indicate that a protein does not contain amino acid residues encoded by vector sequences; that is the native protein contains only those amino acids found in the protein as it occurs in nature. A native protein may be produced by recombinant means or may be isolated from a naturally occurring source.

As used herein the term "portion" when in reference to a protein (as in "a portion of a given protein") refers to fragments of that protein. The fragments may range in size from four consecutive amino acid residues to the entire amino acid sequence minus one amino acid.

5 The term "Southern blot," refers to the analysis of DNA on agarose or acrylamide gels to fractionate the DNA according to size followed by transfer of the DNA from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized DNA is then probed with a labeled probe to detect DNA species complementary to the probe used. The DNA may be cleaved with restriction enzymes prior to electrophoresis.
10 Following electrophoresis, the DNA may be partially depurinated and denatured prior to or during transfer to the solid support. Southern blots are a standard tool of molecular biologists (J. Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press, NY, pp 9.31-9.58 [1989]).

 The term "Northern blot," as used herein refers to the analysis of RNA by
15 electrophoresis of RNA on agarose gels to fractionate the RNA according to size followed by transfer of the RNA from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized RNA is then probed with a labeled probe to detect RNA species complementary to the probe used. Northern blots are a standard tool of molecular biologists (J. Sambrook, *et al.*, *supra*, pp 7.39-7.52 [1989]).

20 The term "Western blot" refers to the analysis of protein(s) (or polypeptides) immobilized onto a support such as nitrocellulose or a membrane. The proteins are run on acrylamide gels to separate the proteins, followed by transfer of the protein from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized proteins are then exposed to antibodies with reactivity against an antigen of interest. The
25 binding of the antibodies may be detected by various methods, including the use of radiolabeled antibodies.

 The term "antigenic determinant" as used herein refers to that portion of an antigen that makes contact with a particular antibody (*i.e.*, an epitope). When a protein or fragment of a protein is used to immunize a host animal, numerous regions of the protein
30 may induce the production of antibodies that bind specifically to a given region or three-dimensional structure on the protein; these regions or structures are referred to as

antigenic determinants. An antigenic determinant may compete with the intact antigen (*i.e.*, the "immunogen" used to elicit the immune response) for binding to an antibody.

The term "transgene" as used herein refers to a foreign, heterologous, or autologous gene that is placed into an organism by introducing the gene into newly fertilized eggs or early embryos. The term "foreign gene" refers to any nucleic acid (*e.g.*, gene sequence) that is introduced into the genome of an animal by experimental manipulations and may include gene sequences found in that animal so long as the introduced gene does not reside in the same location as does the naturally-occurring gene. The term "autologous gene" is intended to encompass variants (*e.g.*, polymorphisms or mutants) of the naturally occurring gene. The term transgene thus encompasses the replacement of the naturally occurring gene with a variant form of the gene.

As used herein, the term "vector" is used in reference to nucleic acid molecules that transfer DNA segment(s) from one cell to another. The term "vehicle" is sometimes used interchangeably with "vector."

The term "expression vector" as used herein refers to a recombinant DNA molecule containing a desired coding sequence and appropriate nucleic acid sequences necessary for the expression of the operably linked coding sequence in a particular host organism. Nucleic acid sequences necessary for expression in prokaryotes usually include a promoter, an operator (optional), and a ribosome binding site, often along with other sequences. Eukaryotic cells are known to utilize promoters, enhancers, and termination and polyadenylation signals.

As used herein, the term "host cell" refers to any eukaryotic or prokaryotic cell (*e.g.*, bacterial cells such as *E. coli*, yeast cells, mammalian cells, avian cells, amphibian cells, plant cells, fish cells, and insect cells), whether located *in vitro* or *in vivo*. For example, host cells may be located in a transgenic animal.

The terms "overexpression" and "overexpressing" and grammatical equivalents, are used in reference to levels of mRNA to indicate a level of expression approximately 3-fold higher than that typically observed in a given tissue in a control or non-transgenic animal. Levels of mRNA are measured using any of a number of techniques known to those skilled in the art including, but not limited to Northern blot analysis (*See*, Example 10, for a protocol for performing Northern blot analysis). Appropriate controls are

included on the Northern blot to control for differences in the amount of RNA loaded from each tissue analyzed (*e.g.*, the amount of 28S rRNA, an abundant RNA transcript present at essentially the same amount in all tissues, present in each sample can be used as a means of normalizing or standardizing the RAD50 mRNA-specific signal observed on Northern blots). The amount of mRNA present in the band corresponding in size to the correctly spliced MCFD2 transgene RNA is quantified; other minor species of RNA which hybridize to the transgene probe are not considered in the quantification of the expression of the transgenic mRNA.

The term "transfection" as used herein refers to the introduction of foreign DNA into eukaryotic cells. Transfection may be accomplished by a variety of means known to the art including calcium phosphate-DNA co-precipitation, DEAE-dextran-mediated transfection, polybrene-mediated transfection, electroporation, microinjection, liposome fusion, lipofection, protoplast fusion, retroviral infection, and biolistics.

The term "stable transfection" or "stably transfected" refers to the introduction and integration of foreign DNA into the genome of the transfected cell. The term "stable transfectant" refers to a cell that has stably integrated foreign DNA into the genomic DNA.

The term "transient transfection" or "transiently transfected" refers to the introduction of foreign DNA into a cell where the foreign DNA fails to integrate into the genome of the transfected cell. The foreign DNA persists in the nucleus of the transfected cell for several days. During this time the foreign DNA is subject to the regulatory controls that govern the expression of endogenous genes in the chromosomes. The term "transient transfectant" refers to cells that have taken up foreign DNA but have failed to integrate this DNA.

The term "calcium phosphate co-precipitation" refers to a technique for the introduction of nucleic acids into a cell. The uptake of nucleic acids by cells is enhanced when the nucleic acid is presented as a calcium phosphate-nucleic acid co-precipitate. The original technique of Graham and van der Eb (Graham and van der Eb, *Virol.*, 52:456 [1973]), has been modified by several groups to optimize conditions for particular types of cells. The art is well aware of these numerous modifications.

A "composition comprising a given polynucleotide sequence" as used herein refers broadly to any composition containing the given polynucleotide sequence. The composition may comprise an aqueous solution. Compositions comprising polynucleotide sequences encoding MCFD2 (*e.g.*, SEQ ID NO:1) or fragments thereof
5 may be employed as hybridization probes. In this case, the MCFD2 encoding polynucleotide sequences are typically employed in an aqueous solution containing salts (*e.g.*, NaCl), detergents (*e.g.*, SDS), and other components (*e.g.*, Denhardt's solution, dry milk, salmon sperm DNA, etc.).

The term "test compound" refers to any chemical entity, pharmaceutical, drug,
10 and the like that can be used to treat or prevent a disease, illness, sickness, or disorder of bodily function, or otherwise alter the physiological or cellular status of a sample. Test compounds comprise both known and potential therapeutic compounds. A test compound can be determined to be therapeutic by screening using the screening methods of the present invention. A "known therapeutic compound" refers to a therapeutic
15 compound that has been shown (*e.g.*, through animal trials or prior experience with administration to humans) to be effective in such treatment or prevention.

The term "sample" as used herein is used in its broadest sense. A sample suspected of containing a human chromosome or sequences associated with a human chromosome may comprise a cell, chromosomes isolated from a cell (*e.g.*, a spread of
20 metaphase chromosomes), genomic DNA (in solution or bound to a solid support such as for Southern blot analysis), RNA (in solution or bound to a solid support such as for Northern blot analysis), cDNA (in solution or bound to a solid support) and the like. A sample suspected of containing a protein may comprise a cell, a portion of a tissue, an extract containing one or more proteins and the like.

25 As used herein, the term "response," when used in reference to an assay, refers to the generation of a detectable signal (*e.g.*, accumulation of reporter protein, increase in ion concentration, accumulation of a detectable chemical product).

As used herein, the term "reporter gene" refers to a gene encoding a protein that may be assayed. Examples of reporter genes include, but are not limited to, luciferase
30 (*See, e.g.*, deWet *et al.*, Mol. Cell. Biol. 7:725 [1987] and U.S. Pat Nos., 6,074,859; 5,976,796; 5,674,713; and 5,618,682; all of which are incorporated herein by reference),

green fluorescent protein (*e.g.*, GenBank Accession Number U43284; a number of GFP variants are commercially available from CLONTECH Laboratories, Palo Alto, CA), chloramphenicol acetyltransferase, β -galactosidase, alkaline phosphatase, and horse radish peroxidase.

5

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to combined deficiency of Factor V and factor VIII (F5F8D), in particular to the MCFD2 protein and nucleic acids encoding MCFD2 protein. The present invention provides compositions and methods for modulating (*i.e.*, increasing
10 or decreasing) the expression of or biological activity of MCFD2. The present invention also provides assays for the detection of *MCFD2* polymorphisms and mutations associated with disease states.

I. MCFD2 Polynucleotides

15 As described above, a novel gene associated with F5F8D has been discovered. Accordingly, the present invention provides nucleic acids encoding *MCFD2* genes, homologs, variants (*e.g.*, polymorphisms and mutants), including but not limited to, those described in SEQ ID NO: 1. In some embodiments, the present invention provides polynucleotide sequences that are capable of hybridizing to SEQ ID NO: 1 under
20 conditions of low to high stringency as long as the polynucleotide sequence capable of hybridizing encodes a protein that retains a biological activity of the naturally occurring MCFD2. In some embodiments, the protein that retains a biological activity of naturally occurring MCFD2 is 70% homologous to wild-type MCFD2, preferably 80% homologous to wild-type MCFD2, more preferably 90% homologous to wild-type
25 MCFD2, and most preferably 95% homologous to wild-type MCFD2. In preferred embodiments, hybridization conditions are based on the melting temperature (T_m) of the nucleic acid binding complex and confer a defined "stringency" as explained above (*See e.g.*, Wahl, *et al.*, Meth. Enzymol., 152:399-407 [1987], incorporated herein by reference).

30 In other embodiments of the present invention, additional alleles of MCFD2 are provided. In preferred embodiments, alleles result from a polymorphism or mutation

(*i.e.*, a change in the nucleic acid sequence) and generally produce altered mRNAs or polypeptides whose structure or function may or may not be altered. Any given gene may have none, one or many allelic forms. Common mutational changes that give rise to alleles are generally ascribed to deletions, additions or substitutions of nucleic acids.

5 Each of these types of changes may occur alone, or in combination with the others, and at the rate of one or more times in a given sequence. Examples of the alleles of the present invention include those encoded by SEQ ID NOs:1 (wild type) and disease alleles described herein (*e.g.*, SEQ ID NOs: 5, 7, 9, 11, 13, 15, and 17).

10 In still other embodiments of the present invention, the nucleotide sequences of the present invention may be engineered in order to alter an MCFD2 coding sequence for a variety of reasons, including but not limited to, alterations which modify the cloning, processing and/or expression of the gene product. For example, mutations may be introduced using techniques that are well known in the art (*e.g.*, site-directed mutagenesis to insert new restriction sites, to alter glycosylation patterns, to change codon preference,

15 etc.).

In some embodiments of the present invention, the polynucleotide sequence of MCFD2 may be extended utilizing the nucleotide sequence (*e.g.*, SEQ ID NO: 1) in various methods known in the art to detect upstream sequences such as promoters and regulatory elements. For example, it is contemplated that restriction-site polymerase

20 chain reaction (PCR) will find use in the present invention. This is a direct method that uses universal primers to retrieve unknown sequence adjacent to a known locus (Gobinda *et al.*, PCR Methods Applic., 2:318-22 [1993]). First, genomic DNA is amplified in the presence of a primer to a linker sequence and a primer specific to the known region. The amplified sequences are then subjected to a second round of PCR

25 with the same linker primer and another specific primer internal to the first one. Products of each round of PCR are transcribed with an appropriate RNA polymerase and sequenced using reverse transcriptase.

In another embodiment, inverse PCR can be used to amplify or extend sequences using divergent primers based on a known region (Triglia *et al.*, Nucleic Acids Res.,

30 16:8186 [1988]). The primers may be designed using Oligo 4.0 (National Biosciences Inc, Plymouth Minn.), or another appropriate program, to be 22-30 nucleotides in length,

to have a GC content of 50% or more, and to anneal to the target sequence at temperatures about 68-72°C. The method uses several restriction enzymes to generate a suitable fragment in the known region of a gene. The fragment is then circularized by intramolecular ligation and used as a PCR template. In still other embodiments, walking PCR is utilized. Walking PCR is a method for targeted gene walking that permits retrieval of unknown sequence (Parker *et al.*, Nucleic Acids Res., 19:3055-60 [1991]). The PROMOTERFINDER kit (Clontech) uses PCR, nested primers and special libraries to "walk in" genomic DNA. This process avoids the need to screen libraries and is useful in finding intron/exon junctions.

Preferred libraries for screening for full-length cDNAs include mammalian libraries that have been size-selected to include larger cDNAs. Also, random primed libraries are preferred, in that they will contain more sequences that contain the 5' and upstream gene regions. A randomly primed library may be particularly useful in case where an oligo d(T) library does not yield full-length cDNA. Genomic mammalian libraries are useful for obtaining introns and extending 5' sequence.

In other embodiments of the present invention, variants of the disclosed MCFD2 sequences are provided. In preferred embodiments, variants result from polymorphisms or mutations (*i.e.*, a change in the nucleic acid sequence) and generally produce altered mRNAs or polypeptides whose structure or function may or may not be altered. Any given gene may have none, one, or many variant forms. Common mutational changes that give rise to variants are generally ascribed to deletions, additions or substitutions of nucleic acids. Each of these types of changes may occur alone, or in combination with the others, and at the rate of one or more times in a given sequence.

It is contemplated that it is possible to modify the structure of a peptide having a function (*e.g.*, MCFD2 function) for such purposes as altering the biological activity (*e.g.*, prevention of F5F8D). Such modified peptides are considered functional equivalents of peptides having an activity of MCFD2 as defined herein. A modified peptide can be produced in which the nucleotide sequence encoding the polypeptide has been altered, such as by substitution, deletion, or addition. In particularly preferred embodiments, these modifications do not significantly reduce the biological activity of the modified MCFD2. In other words, construct "X" can be evaluated in order to

determine whether it is a member of the genus of modified or variant MCFD2's of the present invention as defined functionally, rather than structurally. In preferred embodiments, the activity of variant MCFD24 polypeptides is evaluated by methods described herein (*e.g.*, the generation of transgenic animals).

Moreover, as described above, variant forms of MCFD2 are also contemplated as being equivalent to those peptides and DNA molecules that are set forth in more detail herein. For example, it is contemplated that isolated replacement of a leucine with an isoleucine or valine, an aspartate with a glutamate, a threonine with a serine, or a similar replacement of an amino acid with a structurally related amino acid (*i.e.*, conservative mutations) will not have a major effect on the biological activity of the resulting molecule. Accordingly, some embodiments of the present invention provide variants of MCFD2 disclosed herein containing conservative replacements. Conservative replacements are those that take place within a family of amino acids that are related in their side chains. Genetically encoded amino acids can be divided into four families: (1) acidic (aspartate, glutamate); (2) basic (lysine, arginine, histidine); (3) nonpolar (alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan); and (4) uncharged polar (glycine, asparagine, glutamine, cysteine, serine, threonine, tyrosine). Phenylalanine, tryptophan, and tyrosine are sometimes classified jointly as aromatic amino acids. In similar fashion, the amino acid repertoire can be grouped as (1) acidic (aspartate, glutamate); (2) basic (lysine, arginine, histidine), (3) aliphatic (glycine, alanine, valine, leucine, isoleucine, serine, threonine), with serine and threonine optionally be grouped separately as aliphatic-hydroxyl; (4) aromatic (phenylalanine, tyrosine, tryptophan); (5) amide (asparagine, glutamine); and (6) sulfur -containing (cysteine and methionine) (*e.g.*, Stryer ed., *Biochemistry*, pg. 17-21, 2nd ed, WH Freeman and Co., 1981). Whether a change in the amino acid sequence of a peptide results in a functional polypeptide can be readily determined by assessing the ability of the variant peptide to function in a fashion similar to the wild-type protein. Peptides having more than one replacement can readily be tested in the same manner.

More rarely, a variant includes "nonconservative" changes (*e.g.*, replacement of a glycine with a tryptophan). Analogous minor variations can also include amino acid deletions or insertions, or both. Guidance in determining which amino acid residues can

be substituted, inserted, or deleted without abolishing biological activity can be found using computer programs (*e.g.*, LASERGENE software, DNASTAR Inc., Madison, Wis.).

As described in more detail below, variants may be produced by methods such as directed evolution or other techniques for producing combinatorial libraries of variants, described in more detail below. In still other embodiments of the present invention, the nucleotide sequences of the present invention may be engineered in order to alter a MCFD2 coding sequence including, but not limited to, alterations that modify the cloning, processing, localization, secretion, and/or expression of the gene product. For example, mutations may be introduced using techniques that are well known in the art (*e.g.*, site-directed mutagenesis to insert new restriction sites, alter glycosylation patterns, or change codon preference, etc.).

II. MCFD2 Polypeptides

In other embodiments, the present invention provides MCFD2 polynucleotide sequences that encode MCFD2 polypeptide sequences (*e.g.*, MCFD2 polypeptides; SEQ ID NOs: 2, 6, 8, 10, 12, 14, 16, and 18). Other embodiments of the present invention provide fragments, fusion proteins or functional equivalents of these MCFD2 proteins. In some embodiments, the present invention provides truncation mutants of MCFD2. In still other embodiment of the present invention, nucleic acid sequences corresponding to MCFD2 variants, homologs, and mutants may be used to generate recombinant DNA molecules that direct the expression of the MCFD2 variants, homologs, and mutants in appropriate host cells. In some embodiments of the present invention, the polypeptide may be a naturally purified product, in other embodiments it may be a product of chemical synthetic procedures, and in still other embodiments it may be produced by recombinant techniques using a prokaryotic or eukaryotic host (*e.g.*, by bacterial, yeast, higher plant, insect and mammalian cells in culture). In some embodiments, depending upon the host employed in a recombinant production procedure, the polypeptide of the present invention may be glycosylated or may be non-glycosylated. In other embodiments, the polypeptides of the invention may also include an initial methionine amino acid residue.

In one embodiment of the present invention, due to the inherent degeneracy of the genetic code, DNA sequences other than the polynucleotide sequences of SEQ ID NO:1 that encode substantially the same or a functionally equivalent amino acid sequence, may be used to clone and express MCFD2. In general, such polynucleotide sequences
5 hybridize to SEQ ID NO:1 under conditions of high to medium stringency as described above. As will be understood by those of skill in the art, it may be advantageous to produce MCFD2-encoding nucleotide sequences possessing non-naturally occurring codons. Therefore, in some preferred embodiments, codons preferred by a particular prokaryotic or eukaryotic host (Murray *et al.*, Nucl. Acids Res., 17 [1989]) are selected,
10 for example, to increase the rate of MCFD2 expression or to produce recombinant RNA transcripts having desirable properties, such as a longer half-life, than transcripts produced from naturally occurring sequence.

1. Vectors for Production of MCFD2

15 The polynucleotides of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may be included in any one of a variety of expression vectors for expressing a polypeptide. In some embodiments of the present invention, vectors include, but are not limited to, chromosomal, nonchromosomal and synthetic DNA sequences (*e.g.*, derivatives of SV40,
20 bacterial plasmids, phage DNA; baculovirus, yeast plasmids, vectors derived from combinations of plasmids and phage DNA, and viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies). It is contemplated that any vector may be used as long as it is replicable and viable in the host.

In particular, some embodiments of the present invention provide recombinant
25 constructs comprising one or more of the sequences as broadly described above (*e.g.*, SEQ ID NOs: 1, 5, 7, 9, 11, 13, 15, and 17). In some embodiments of the present invention, the constructs comprise a vector, such as a plasmid or viral vector, into which a sequence of the invention has been inserted, in a forward or reverse orientation. In still other embodiments, the heterologous structural sequence (*e.g.*, SEQ ID NO:1) is
30 assembled in appropriate phase with translation initiation and termination sequences. In preferred embodiments of the present invention, the appropriate DNA sequence is

inserted into the vector using any of a variety of procedures. In general, the DNA sequence is inserted into an appropriate restriction endonuclease site(s) by procedures known in the art.

Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. Such vectors include, but are not limited to, the following vectors: 1) Bacterial -- pQE70, pQE60, pQE-9 (Qiagen), pBS, pD10, phagescript, psiX174, pbluescript SK, pBSKS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); 2) Eukaryotic -- pWLNEO, pSV2CAT, pOG44, PXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia); and 3) Baculovirus – pPbac and pMbac (Stratagene). Any other plasmid or vector may be used as long as they are replicable and viable in the host. In some preferred embodiments of the present invention, mammalian expression vectors comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation sites, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking non-transcribed sequences. In other embodiments, DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required non-transcribed genetic elements.

In certain embodiments of the present invention, the DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. Promoters useful in the present invention include, but are not limited to, the LTR or SV40 promoter, the *E. coli lac* or *trp*, the phage lambda P_L and P_R, T3 and T7 promoters, and the cytomegalovirus (CMV) immediate early, herpes simplex virus (HSV) thymidine kinase, and mouse metallothionein-I promoters and other promoters known to control expression of gene in prokaryotic or eukaryotic cells or their viruses. In other embodiments of the present invention, recombinant expression vectors include origins of replication and selectable markers permitting transformation of the host cell (e.g., dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or tetracycline or ampicillin resistance in *E. coli*).

In some embodiments of the present invention, transcription of the DNA encoding the polypeptides of the present invention by higher eukaryotes is increased by inserting an enhancer sequence into the vector. Enhancers are *cis*-acting elements of DNA, usually

about from 10 to 300 bp that act on a promoter to increase its transcription. Enhancers useful in the present invention include, but are not limited to, the SV40 enhancer on the late side of the replication origin bp 100 to 270, a cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and adenovirus enhancers.

In other embodiments, the expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. In still other embodiments of the present invention, the vector may also include appropriate sequences for amplifying expression.

2. Host Cells for Production of MCFD2

In a further embodiment, the present invention provides host cells containing the above-described constructs. In some embodiments of the present invention, the host cell is a higher eukaryotic cell (*e.g.*, a mammalian or insect cell). In other embodiments of the present invention, the host cell is a lower eukaryotic cell (*e.g.*, a yeast cell). In still other embodiments of the present invention, the host cell can be a prokaryotic cell (*e.g.*, a bacterial cell). Specific examples of host cells include, but are not limited to, *Escherichia coli*, *Salmonella typhimurium*, *Bacillus subtilis*, and various species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*, as well as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila* S2 cells, *Spodoptera* Sf9 cells, Chinese hamster ovary (CHO) cells, COS-7 lines of monkey kidney fibroblasts, (Gluzman, Cell 23:175 [1981]), C127, 3T3, 293, 293T, HeLa and BHK cell lines.

The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. In some embodiments, introduction of the construct into the host cell can be accomplished by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (*See e.g.*, Davis *et al.*, Basic Methods in Molecular Biology, [1986]). Alternatively, in some embodiments of the present invention, the polypeptides of the invention can be synthetically produced by conventional peptide synthesizers.

Proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed

to produce such proteins using RNAs derived from the DNA constructs of the present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., [1989].

5 In some embodiments of the present invention, following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is induced by appropriate means (*e.g.*, temperature shift or chemical induction) and cells are cultured for an additional period. In other embodiments of the present invention, cells are typically harvested by centrifugation, disrupted by physical or
10 chemical means, and the resulting crude extract retained for further purification. In still other embodiments of the present invention, microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents.

15 3. Purification of MCFD2

The present invention also provides methods for recovering and purifying MCFD2 from recombinant cell cultures including, but not limited to, ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity
20 chromatography, hydroxylapatite chromatography and lectin chromatography. In other embodiments of the present invention, protein-refolding steps can be used as necessary, in completing configuration of the mature protein. In still other embodiments of the present invention, high performance liquid chromatography (HPLC) can be employed for final purification steps.

25 The present invention further provides polynucleotides having the coding sequence (*e.g.*, SEQ ID NO: 1) fused in frame to a marker sequence that allows for purification of the polypeptide of the present invention. A non-limiting example of a marker sequence is a hexahistidine tag which may be supplied by a vector, preferably a pQE-9 vector, which provides for purification of the polypeptide fused to the marker in
30 the case of a bacterial host, or, for example, the marker sequence may be a hemagglutinin (HA) tag when a mammalian host (*e.g.*, COS-7 cells) is used. The HA tag corresponds to

an epitope derived from the influenza hemagglutinin protein (Wilson *et al.*, Cell, 37:767 [1984]).

4. Truncation Mutants of MCFD2

5 In addition, the present invention provides fragments of MCFD2. As described above, truncations of MCFD2 were found in families with F5F8D. In some embodiments of the present invention, when expression of a portion of the MCFD2 protein is desired, it may be necessary to add a start codon (ATG) to the oligonucleotide fragment containing the desired sequence to be expressed. It is well known in the art that a methionine at the
10 N-terminal position can be enzymatically cleaved by the use of the enzyme methionine aminopeptidase (MAP). MAP has been cloned from *E. coli* (Ben-Bassat *et al.*, J. Bacteriol., 169:751 [1987]) and *Salmonella typhimurium* and its *in vitro* activity has been demonstrated on recombinant proteins (Miller *et al.*, Proc. Natl. Acad. Sci. USA 84:2718 [1990]). Therefore, removal of an N-terminal methionine, if desired, can be achieved
15 either *in vivo* by expressing such recombinant polypeptides in a host which produces MAP (*e.g.*, *E. coli* or CM89 or *S. cerevisiae*), or *in vitro* by use of purified MAP.

5. Fusion Proteins Containing MCFD2

The present invention also provides fusion proteins incorporating all or part of
20 MCFD2. Accordingly, in some embodiments of the present invention, the coding sequences for the polypeptide can be incorporated as a part of a fusion gene including a nucleotide sequence encoding a different polypeptide. It is contemplated that this type of expression system will find use under conditions where it is desirable to produce an immunogenic fragment of a MCFD2 protein. In some embodiments of the present
25 invention, the VP6 capsid protein of rotavirus is used as an immunologic carrier protein for portions of the MCFD2 polypeptide, either in the monomeric form or in the form of a viral particle. In other embodiments of the present invention, the nucleic acid sequences corresponding to the portion of MCFD2 against which antibodies are to be raised can be incorporated into a fusion gene construct which includes coding sequences for a late
30 vaccinia virus structural protein to produce a set of recombinant viruses expressing fusion proteins comprising a portion of MCFD2 as part of the virion. It has been demonstrated

with the use of immunogenic fusion proteins utilizing the hepatitis B surface antigen fusion proteins that recombinant hepatitis B virions can be utilized in this role as well. Similarly, in other embodiments of the present invention, chimeric constructs coding for fusion proteins containing a portion of MCFD2 and the poliovirus capsid protein are
5 created to enhance immunogenicity of the set of polypeptide antigens (*See e.g.*, EP Publication No. 025949; and Evans *et al.*, Nature 339:385 [1989]; Huang *et al.*, J. Virol., 62:3855 [1988]; and Schlienger *et al.*, J. Virol., 66:2 [1992]).

In still other embodiments of the present invention, the multiple antigen peptide system for peptide-based immunization can be utilized. In this system, a desired portion
10 of MCFD2 is obtained directly from organo-chemical synthesis of the peptide onto an oligomeric branching lysine core (*see e.g.*, Posnett *et al.*, J. Biol. Chem., 263:1719 [1988]; and Nardelli *et al.*, J. Immunol., 148:914 [1992]). In other embodiments of the present invention, antigenic determinants of the MCFD2 proteins can also be expressed and presented by bacterial cells.

In addition to utilizing fusion proteins to enhance immunogenicity, it is widely
15 appreciated that fusion proteins can also facilitate the expression of proteins, such as the MCFD2 protein of the present invention. Accordingly, in some embodiments of the present invention, MCFD2 can be generated as a glutathione-S-transferase (*i.e.*, GST fusion protein). It is contemplated that such GST fusion proteins will enable easy
20 purification of MCFD2, such as by the use of glutathione-derivatized matrices (*See e.g.*, Ausabel *et al.* (eds.), Current Protocols in Molecular Biology, John Wiley & Sons, NY [1991]). In another embodiment of the present invention, a fusion gene coding for a purification leader sequence, such as a poly-(His)/enterokinase cleavage site sequence at the N-terminus of the desired portion of MCFD2, can allow purification of the expressed
25 MCFD2 fusion protein by affinity chromatography using a Ni²⁺ metal resin. In still another embodiment of the present invention, the purification leader sequence can then be subsequently removed by treatment with enterokinase (*See e.g.*, Hochuli *et al.*, J. Chromatogr., 411:177 [1987]; and Janknecht *et al.*, Proc. Natl. Acad. Sci. USA 88:8972).

Techniques for making fusion genes are well known. Essentially, the joining of
30 various DNA fragments coding for different polypeptide sequences is performed in accordance with conventional techniques, employing blunt-ended or stagger-ended

termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and enzymatic ligation. In another embodiment of the present invention, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, in other embodiments of the present invention, PCR amplification of gene fragments can be carried out using anchor primers which give rise to complementary overhangs between two consecutive gene fragments which can subsequently be annealed to generate a chimeric gene sequence (*See e.g.*, Current Protocols in Molecular Biology, *supra*).

6. Variants of MCFD2

Still other embodiments of the present invention provide mutant or variant forms of MCFD2 (*i.e.*, muteins). It is possible to modify the structure of a peptide having an activity of MCFD2 for such purposes as enhancing therapeutic or prophylactic efficacy, or stability (*e.g.*, *ex vivo* shelf life, and/or resistance to proteolytic degradation *in vivo*). Such modified peptides are considered functional equivalents of peptides having an activity of the subject MCFD2 proteins as defined herein. A modified peptide can be produced in which the amino acid sequence has been altered, such as by amino acid substitution, deletion, or addition.

Moreover, as described above, variant forms (*e.g.*, mutants or polymorphic sequences) of the subject MCFD2 proteins are also contemplated as being equivalent to those peptides and DNA molecules that are set forth in more detail. For example, as described above, the present invention encompasses mutant and variant proteins that contain conservative or non-conservative amino acid substitutions.

This invention further contemplates a method of generating sets of combinatorial mutants of the present MCFD2 proteins, as well as truncation mutants, and is especially useful for identifying potential variant sequences (*i.e.*, mutants or polymorphic sequences) that are involved in hematologic disease or resistance to hematologic disease. The purpose of screening such combinatorial libraries is to generate, for example, novel MCFD2 variants that can act as either agonists or antagonists, or alternatively, possess novel activities all together.

Therefore, in some embodiments of the present invention, MCFD2 variants are engineered by the present method to provide altered (*e.g.*, increased or decreased) biological activity. In other embodiments of the present invention, combinatorially-derived variants are generated which have a selective potency relative to a naturally occurring MCFD2. Such proteins, when expressed from recombinant DNA constructs, can be used in gene therapy protocols.

Still other embodiments of the present invention provide MCFD2 variants that have intracellular half-lives dramatically different than the corresponding wild-type protein. For example, the altered protein can be rendered either more stable or less stable to proteolytic degradation or other cellular process that result in destruction of, or otherwise inactivate MCFD2. Such variants, and the genes which encode them, can be utilized to alter the location of MCFD2 expression by modulating the half-life of the protein. For instance, a short half-life can give rise to more transient MCFD2 biological effects and, when part of an inducible expression system, can allow tighter control of MCFD2 levels within the cell. As above, such proteins, and particularly their recombinant nucleic acid constructs, can be used in gene therapy protocols.

In still other embodiments of the present invention, MCFD2 variants are generated by the combinatorial approach to act as antagonists, in that they are able to interfere with the ability of the corresponding wild-type protein to regulate cell function.

In some embodiments of the combinatorial mutagenesis approach of the present invention, the amino acid sequences for a population of MCFD2 homologs, variants or other related proteins are aligned, preferably to promote the highest homology possible. Such a population of variants can include, for example, MCFD2 homologs from one or more species, or MCFD2 variants from the same species but which differ due to mutation or polymorphisms. Amino acids that appear at each position of the aligned sequences are selected to create a degenerate set of combinatorial sequences.

In a preferred embodiment of the present invention, the combinatorial MCFD2 library is produced by way of a degenerate library of genes encoding a library of polypeptides which each include at least a portion of potential MCFD2 protein sequences. For example, a mixture of synthetic oligonucleotides can be enzymatically ligated into gene sequences such that the degenerate set of potential MCFD2 sequences

are expressible as individual polypeptides, or alternatively, as a set of larger fusion proteins (*e.g.*, for phage display) containing the set of MCFD2 sequences therein.

There are many ways by which the library of potential MCFD2 homologs and variants can be generated from a degenerate oligonucleotide sequence. In some
5 embodiments, chemical synthesis of a degenerate gene sequence is carried out in an automatic DNA synthesizer, and the synthetic genes are ligated into an appropriate gene for expression. The purpose of a degenerate set of genes is to provide, in one mixture, all of the sequences encoding the desired set of potential MCFD2 sequences. The synthesis of degenerate oligonucleotides is well known in the art (*See e.g.*, Narang, Tetrahedron
10 Lett., 39:39 [1983]; Itakura *et al.*, Recombinant DNA, in Walton (ed.), *Proceedings of the 3rd Cleveland Symposium on Macromolecules*, Elsevier, Amsterdam, pp 273-289 [1981]; Itakura *et al.*, Annu. Rev. Biochem., 53:323 [1984]; Itakura *et al.*, Science 198:1056 [1984]; Ike *et al.*, Nucl. Acid Res., 11:477 [1983]). Such techniques have been employed in the directed evolution of other proteins (*See e.g.*, Scott *et al.*, Science
15 249:386 [1980]; Roberts *et al.*, Proc. Natl. Acad. Sci. USA 89:2429 [1992]; Devlin *et al.*, Science 249: 404 [1990]; Cwirla *et al.*, Proc. Natl. Acad. Sci. USA 87: 6378 [1990]; each of which is herein incorporated by reference; as well as U.S. Pat. Nos. 5,223,409, 5,198,346, and 5,096,815; each of which is incorporated herein by reference).

It is contemplated that the MCFD2 nucleic acids (*e.g.*, SEQ ID NO:1, and
20 fragments and variants thereof) can be utilized as starting nucleic acids for directed evolution. These techniques can be utilized to develop MCFD2 variants having desirable properties such as increased or decreased biological activity.

In some embodiments, artificial evolution is performed by random mutagenesis (*e.g.*, by utilizing error-prone PCR to introduce random mutations into a given coding
25 sequence). This method requires that the frequency of mutation be finely tuned. As a general rule, beneficial mutations are rare, while deleterious mutations are common. This is because the combination of a deleterious mutation and a beneficial mutation often results in an inactive enzyme. The ideal number of base substitutions for targeted gene is usually between 1.5 and 5 (Moore and Arnold, Nat. Biotech., 14, 458 [1996]; Leung *et al.*, Technique, 1:11 [1989]; Eckert and Kunkel, PCR Methods Appl., 1:17-24 [1991];
30 Caldwell and Joyce, PCR Methods Appl., 2:28 [1992]; and Zhao and Arnold, Nuc. Acids.

Res., 25:1307 [1997]). After mutagenesis, the resulting clones are selected for desirable activity (*e.g.*, screened for MCFD2 activity). Successive rounds of mutagenesis and selection are often necessary to develop enzymes with desirable properties. It should be noted that only the useful mutations are carried over to the next round of mutagenesis.

5 In other embodiments of the present invention, the polynucleotides of the present invention are used in gene shuffling or sexual PCR procedures (*e.g.*, Smith, *Nature*, 370:324 [1994]; U.S. Pat. Nos. 5,837,458; 5,830,721; 5,811,238; 5,733,731; all of which are herein incorporated by reference). Gene shuffling involves random fragmentation of several mutant DNAs followed by their reassembly by PCR into full length molecules.

10 Examples of various gene shuffling procedures include, but are not limited to, assembly following DNase treatment, the staggered extension process (STEP), and random priming *in vitro* recombination. In the DNase mediated method, DNA segments isolated from a pool of positive mutants are cleaved into random fragments with DNaseI and subjected to multiple rounds of PCR with no added primer. The lengths of random fragments

15 approach that of the uncleaved segment as the PCR cycles proceed, resulting in mutations in present in different clones becoming mixed and accumulating in some of the resulting sequences. Multiple cycles of selection and shuffling have led to the functional enhancement of several enzymes (Stemmer, *Nature*, 370:398 [1994]; Stemmer, *Proc. Natl. Acad. Sci. USA*, 91:10747 [1994]; Cramer *et al.*, *Nat. Biotech.*, 14:315 [1996];

20 Zhang *et al.*, *Proc. Natl. Acad. Sci. USA*, 94:4504 [1997]; and Cramer *et al.*, *Nat. Biotech.*, 15:436 [1997]). Variants produced by directed evolution can be screened for MCFD2 activity by the methods described herein.

 A wide range of techniques are known in the art for screening gene products of combinatorial libraries made by point mutations, and for screening cDNA libraries for
25 gene products having a certain property. Such techniques will be generally adaptable for rapid screening of the gene libraries generated by the combinatorial mutagenesis or recombination of MCFD2 homologs or variants. The most widely used techniques for screening large gene libraries typically comprises cloning the gene library into replicable expression vectors, transforming appropriate cells with the resulting library of vectors,
30 and expressing the combinatorial genes under conditions in which detection of a desired

activity facilitates relatively easy isolation of the vector encoding the gene whose product was detected.

7. Chemical Synthesis of MCFD2

5 In an alternate embodiment of the invention, the coding sequence of MCFD2 is synthesized, whole or in part, using chemical methods well known in the art (*See e.g.*, Caruthers *et al.*, Nucl. Acids Res. Symp. Ser., 7:215 [1980]; Crea and Horn, Nucl. Acids Res., 9:2331 [1980]; Matteucci and Caruthers, Tetrahedron Lett., 21:719 [1980]; and Chow and Kempe, Nucl. Acids Res., 9:2807 [1981]). In other embodiments of the
10 present invention, the protein itself is produced using chemical methods to synthesize either an entire MCFD2 amino acid sequence or a portion thereof. For example, peptides can be synthesized by solid phase techniques, cleaved from the resin, and purified by preparative high performance liquid chromatography (*See e.g.*, Creighton, *Proteins Structures And Molecular Principles*, W H Freeman and Co, New York N.Y. [1983]). In
15 other embodiments of the present invention, the composition of the synthetic peptides is confirmed by amino acid analysis or sequencing (*See e.g.*, Creighton, *supra*).

Direct peptide synthesis can be performed using various solid-phase techniques (Roberge *et al.*, Science 269:202 [1995]) and automated synthesis may be achieved, for example, using ABI 431A Peptide Synthesizer (Perkin Elmer) in accordance with the
20 instructions provided by the manufacturer. Additionally, the amino acid sequence of MCFD2, or any part thereof, may be altered during direct synthesis and/or combined using chemical methods with other sequences to produce a variant polypeptide.

III. Detection of MCFD2 Alleles

25 In some embodiments, the present invention provides methods of detecting the presence of wild-type or variant (*e.g.*, mutant or polymorphic) MCFD2 nucleic acids or polypeptides. The detection of mutant MCFD2 polypeptides finds use in the diagnosis of disease (*e.g.*, F5F8D).

A. MCFD2 Alleles

In some embodiments, the present invention includes alleles of MCFD2 that increase a patient's susceptibility to F5F8D (*e.g.*, including, but not limited to, SEQ ID NOs: 5, 7, 9, 11, 13, 15, and 17). However, the present invention is not limited to the mutation described in SEQ ID NOs: 5, 7, 9, 11, 13, 15, and 17. Any mutation that results in the undesired phenotype (*e.g.*, coagulopathy) is within the scope of the present invention.

B. Detection of MCFD2 Alleles

Accordingly, the present invention provides methods for determining whether a patient has an increased susceptibility for F5F8D by determining whether the individual has a variant MCFD2 allele. In other embodiments, the present invention provides methods for providing a prognosis of increased risk for F5F8D to an individual based on the presence or absence of one or more variant alleles of MCFD2. In preferred embodiments, the variation causes a truncation of the MCFD2 protein.

A number of methods are available for analysis of variant (*e.g.*, mutant or polymorphic) nucleic acid sequences. Assays for detection variants (*e.g.*, polymorphisms or mutations) fall into several categories, including, but not limited to direct sequencing assays, fragment polymorphism assays, hybridization assays, and computer based data analysis. Protocols and commercially available kits or services for performing multiple variations of these assays are available. In some embodiments, assays are performed in combination or in hybrid (*e.g.*, different reagents or technologies from several assays are combined to yield one assay). The following assays are useful in the present invention.

1. Direct sequencing Assays

In some embodiments of the present invention, variant sequences are detected using a direct sequencing technique. In these assays, DNA samples are first isolated from a subject using any suitable method. In some embodiments, the region of interest is cloned into a suitable vector and amplified by growth in a host cell (*e.g.*, a bacteria). In other embodiments, DNA in the region of interest is amplified using PCR.

Following amplification, DNA in the region of interest (*e.g.*, the region containing the SNP or mutation of interest) is sequenced using any suitable method, including but

not limited to manual sequencing using radioactive marker nucleotides, or automated sequencing. The results of the sequencing are displayed using any suitable method. The sequence is examined and the presence or absence of a given SNP or mutation is determined.

5

2. PCR Assay

In some embodiments of the present invention, variant sequences are detected using a PCR-based assay. In some embodiments, the PCR assay comprises the use of oligonucleotide primers that hybridize only to the variant or wild type allele of MCFD2 (e.g., to the region of polymorphism or mutation). Both sets of primers are used to
10 amplify a sample of DNA. If only the mutant primers result in a PCR product, then the patient has the mutant MCFD2 allele. If only the wild-type primers result in a PCR product, then the patient has the wild type allele of MCFD2.

15

3. Mutational detection by dHPLC

In some embodiments of the present invention, variant sequences are detected using a PCR-based assay with consecutive detection of nucleotide variants by dHPLC (denaturing high performance liquid chromatography). Exemplary systems and Methods for dHPLC include, but are not limited to, WAVE (Transgenomic, Inc; Omaha,
20 NE) or VARIAN equipment (Palo Alto, CA).

4. Fragment Length Polymorphism Assays

In some embodiments of the present invention, variant sequences are detected using a fragment length polymorphism assay. In a fragment length polymorphism assay,
25 a unique DNA banding pattern based on cleaving the DNA at a series of positions is generated using an enzyme (e.g., a restriction enzyme or a CLEAVASE I [Third Wave Technologies, Madison, WI] enzyme). DNA fragments from a sample containing a SNP or a mutation will have a different banding pattern than wild type.

30

a. RFLP Assay

In some embodiments of the present invention, variant sequences are detected using a restriction fragment length polymorphism assay (RFLP). The region of interest is first isolated using PCR. The PCR products are then cleaved with restriction enzymes known to give a unique length fragment for a given polymorphism. The restriction-enzyme digested PCR products are separated by agarose gel electrophoresis and visualized by ethidium bromide staining. The length of the fragments is compared to molecular weight markers and fragments generated from wild-type and mutant controls.

b. CFLP Assay

In other embodiments, variant sequences are detected using a CLEAVASE fragment length polymorphism assay (CFLP; Third Wave Technologies, Madison, WI; *See e.g.*, U.S. Patent Nos. 5,843,654; 5,843,669; 5,719,208; and 5,888,780; each of which is herein incorporated by reference). This assay is based on the observation that when single strands of DNA fold on themselves, they assume higher order structures that are highly individual to the precise sequence of the DNA molecule. These secondary structures involve partially duplexed regions of DNA such that single stranded regions are juxtaposed with double stranded DNA hairpins. The CLEAVASE I enzyme, is a structure-specific, thermostable nuclease that recognizes and cleaves the junctions between these single-stranded and double-stranded regions.

The region of interest is first isolated, for example, using PCR. Then, DNA strands are separated by heating. Next, the reactions are cooled to allow intrastrand secondary structure to form. The PCR products are then treated with the CLEAVASE I enzyme to generate a series of fragments that are unique to a given SNP or mutation. The CLEAVASE enzyme treated PCR products are separated and detected (*e.g.*, by agarose gel electrophoresis) and visualized (*e.g.*, by ethidium bromide staining). The length of the fragments is compared to molecular weight markers and fragments generated from wild-type and mutant controls.

5. Hybridization Assays

In preferred embodiments of the present invention, variant sequences are detected a hybridization assay. In a hybridization assay, the presence of absence of a given SNP

or mutation is determined based on the ability of the DNA from the sample to hybridize to a complementary DNA molecule (*e.g.*, a oligonucleotide probe). A variety of hybridization assays using a variety of technologies for hybridization and detection are available. A description of a selection of assays is provided below.

5

a. Direct Detection of Hybridization

In some embodiments, hybridization of a probe to the sequence of interest (*e.g.*, a SNP or mutation) is detected directly by visualizing a bound probe (*e.g.*, a Northern or Southern assay; *See e.g.*, Ausabel *et al.* (eds.), Current Protocols in Molecular Biology, 10 John Wiley & Sons, NY [1991]). In a these assays, genomic DNA (Southern) or RNA (Northern) is isolated from a subject. The DNA or RNA is then cleaved with a series of restriction enzymes that cleave infrequently in the genome and not near any of the markers being assayed. The DNA or RNA is then separated (*e.g.*, on an agarose gel) and transferred to a membrane. A labeled (*e.g.*, by incorporating a radionucleotide) probe or 15 probes specific for the SNP or mutation being detected is allowed to contact the membrane under a condition or low, medium, or high stringency conditions. Unbound probe is removed and the presence of binding is detected by visualizing the labeled probe.

20 **b. Detection of Hybridization Using "DNA Chip" Assays**

In some embodiments of the present invention, variant sequences are detected using a DNA chip hybridization assay. In this assay, a series of oligonucleotide probes are affixed to a solid support. The oligonucleotide probes are designed to be unique to a given SNP or mutation. The DNA sample of interest is contacted with the DNA "chip" 25 and hybridization is detected.

In some embodiments, the DNA chip assay is a GeneChip (Affymetrix, Santa Clara, CA; *See e.g.*, U.S. Patent Nos. 6,045,996; 5,925,525; and 5,858,659; each of which is herein incorporated by reference) assay. The GeneChip technology uses miniaturized, high-density arrays of oligonucleotide probes affixed to a "chip." Probe arrays are 30 manufactured by Affymetrix's light-directed chemical synthesis process, which combines solid-phase chemical synthesis with photolithographic fabrication techniques employed

in the semiconductor industry. Using a series of photolithographic masks to define chip exposure sites, followed by specific chemical synthesis steps, the process constructs high-density arrays of oligonucleotides, with each probe in a predefined position in the array. Multiple probe arrays are synthesized simultaneously on a large glass wafer. The wafers are then diced, and individual probe arrays are packaged in injection-molded plastic cartridges, which protect them from the environment and serve as chambers for hybridization.

The nucleic acid to be analyzed is isolated, amplified by PCR, and labeled with a fluorescent reporter group. The labeled DNA is then incubated with the array using a fluidics station. The array is then inserted into the scanner, where patterns of hybridization are detected. The hybridization data are collected as light emitted from the fluorescent reporter groups already incorporated into the target, which is bound to the probe array. Probes that perfectly match the target generally produce stronger signals than those that have mismatches. Since the sequence and position of each probe on the array are known, by complementarity, the identity of the target nucleic acid applied to the probe array can be determined.

In other embodiments, a DNA microchip containing electronically captured probes (Nanogen, San Diego, CA) is utilized (*See e.g.*, U.S. Patent Nos. 6,017,696; 6,068,818; and 6,051,380; each of which are herein incorporated by reference). Through the use of microelectronics, Nanogen's technology enables the active movement and concentration of charged molecules to and from designated test sites on its semiconductor microchip. DNA capture probes unique to a given SNP or mutation are electronically placed at, or "addressed" to, specific sites on the microchip. Since DNA has a strong negative charge, it can be electronically moved to an area of positive charge.

First, a test site or a row of test sites on the microchip is electronically activated with a positive charge. Next, a solution containing the DNA probes is introduced onto the microchip. The negatively charged probes rapidly move to the positively charged sites, where they concentrate and are chemically bound to a site on the microchip. The microchip is then washed and another solution of distinct DNA probes is added until the array of specifically bound DNA probes is complete.

A test sample is then analyzed for the presence of target DNA molecules by determining which of the DNA capture probes hybridize, with complementary DNA in the test sample (*e.g.*, a PCR amplified gene of interest). An electronic charge is also used to move and concentrate target molecules to one or more test sites on the microchip. The electronic concentration of sample DNA at each test site promotes rapid hybridization of sample DNA with complementary capture probes (hybridization may occur in minutes). To remove any unbound or nonspecifically bound DNA from each site, the polarity or charge of the site is reversed to negative, thereby forcing any unbound or nonspecifically bound DNA back into solution away from the capture probes. A laser-based fluorescence scanner is used to detect binding,

In still further embodiments, an array technology based upon the segregation of fluids on a flat surface (chip) by differences in surface tension (ProtoGene, Palo Alto, CA) is utilized (*See e.g.*, U.S. Patent Nos. 6,001,311; 5,985,551; and 5,474,796; each of which is herein incorporated by reference). Protogene's technology is based on the fact that fluids can be segregated on a flat surface by differences in surface tension that have been imparted by chemical coatings. Once so segregated, oligonucleotide probes are synthesized directly on the chip by ink-jet printing of reagents. The array with its reaction sites defined by surface tension is mounted on a X/Y translation stage under a set of four piezoelectric nozzles, one for each of the four standard DNA bases. The translation stage moves along each of the rows of the array and the appropriate reagent is delivered to each of the reaction site. For example, the A amidite is delivered only to the sites where amidite A is to be coupled during that synthesis step and so on. Common reagents and washes are delivered by flooding the entire surface and then removing them by spinning.

DNA probes unique for the SNP or mutation of interest are affixed to the chip using Protogene's technology. The chip is then contacted with the PCR-amplified genes of interest. Following hybridization, unbound DNA is removed and hybridization is detected using any suitable method (*e.g.*, by fluorescence de-quenching of an incorporated fluorescent group).

In yet other embodiments, a "bead array" is used for the detection of polymorphisms (Illumina, San Diego, CA; *See e.g.*, PCT Publications WO 99/67641 and

WO 00/39587, each of which is herein incorporated by reference). Illumina uses a BEAD ARRAY technology that combines fiber optic bundles and beads that self-assemble into an array. Each fiber optic bundle contains thousands to millions of individual fibers depending on the diameter of the bundle. The beads are coated with an oligonucleotide specific for the detection of a given SNP or mutation. Batches of beads are combined to form a pool specific to the array. To perform an assay, the BEAD ARRAY is contacted with a prepared subject sample (*e.g.*, DNA). Hybridization is detected using any suitable method.

c. Enzymatic Detection of Hybridization

In some embodiments of the present invention, hybridization is detected by enzymatic cleavage of specific structures (INVADER assay, Third Wave Technologies; *See e.g.*, U.S. Patent Nos. 5,846,717, 6,090,543; 6,001,567; 5,985,557; and 5,994,069; each of which is herein incorporated by reference). The INVADER assay detects specific DNA and RNA sequences by using structure-specific enzymes to cleave a complex formed by the hybridization of overlapping oligonucleotide probes. Elevated temperature and an excess of one of the probes enable multiple probes to be cleaved for each target sequence present without temperature cycling. These cleaved probes then direct cleavage of a second labeled probe. The secondary probe oligonucleotide can be 5'-end labeled with fluorescein that is quenched by an internal dye. Upon cleavage, the de-quenched fluorescein labeled product may be detected using a standard fluorescence plate reader.

The INVADER assay detects specific mutations and SNPs in unamplified genomic DNA. The isolated DNA sample is contacted with the first probe specific either for a SNP/mutation or wild type sequence and allowed to hybridize. Then a secondary probe, specific to the first probe, and containing the fluorescein label, is hybridized and the enzyme is added. Binding is detected by using a fluorescent plate reader and comparing the signal of the test sample to known positive and negative controls.

In some embodiments, hybridization of a bound probe is detected using a TaqMan assay (PE Biosystems, Foster City, CA; *See e.g.*, U.S. Patent Nos. 5,962,233 and 5,538,848, each of which is herein incorporated by reference). The assay is performed during a PCR reaction. The TaqMan assay exploits the 5'-3' exonuclease activity of the

AMPLITAQ GOLD DNA polymerase. A probe, specific for a given allele or mutation, is included in the PCR reaction. The probe consists of an oligonucleotide with a 5'-reporter dye (*e.g.*, a fluorescent dye) and a 3'-quencher dye. During PCR, if the probe is bound to its target, the 5'-3' nucleolytic activity of the AMPLITAQ GOLD polymerase cleaves the probe between the reporter and the quencher dye. The separation of the reporter dye from the quencher dye results in an increase of fluorescence. The signal accumulates with each cycle of PCR and can be monitored with a fluorimeter.

In still further embodiments, polymorphisms are detected using the SNP-IT primer extension assay (Orchid Biosciences, Princeton, NJ; *See e.g.*, U.S. Patent Nos. 5,952,174 and 5,919,626, each of which is herein incorporated by reference). In this assay, SNPs are identified by using a specially synthesized DNA primer and a DNA polymerase to selectively extend the DNA chain by one base at the suspected SNP location. DNA in the region of interest is amplified and denatured. Polymerase reactions are then performed using miniaturized systems called microfluidics. Detection is accomplished by adding a label to the nucleotide suspected of being at the SNP or mutation location. Incorporation of the label into the DNA can be detected by any suitable method (*e.g.*, if the nucleotide contains a biotin label, detection is via a fluorescently labeled antibody specific for biotin).

6. Mass Spectroscopy Assay

In some embodiments, a MassARRAY system (Sequenom, San Diego, CA.) is used to detect variant sequences (*See e.g.*, U.S. Patent Nos. 6,043,031; 5,777,324; and 5,605,798; each of which is herein incorporated by reference). DNA is isolated from blood samples using standard procedures. Next, specific DNA regions containing the mutation or SNP of interest, about 200 base pairs in length, are amplified by PCR. The amplified fragments are then attached by one strand to a solid surface and the non-immobilized strands are removed by standard denaturation and washing. The remaining immobilized single strand then serves as a template for automated enzymatic reactions that produce genotype specific diagnostic products.

Very small quantities of the enzymatic products, typically five to ten nanoliters, are then transferred to a SpectroCHIP array for subsequent automated analysis with the

SpectroREADER mass spectrometer. Each spot is preloaded with light absorbing crystals that form a matrix with the dispensed diagnostic product. The MassARRAY system uses MALDI-TOF (Matrix Assisted Laser Desorption Ionization - Time of Flight) mass spectrometry. In a process known as desorption, the matrix is hit with a pulse from a laser beam. Energy from the laser beam is transferred to the matrix and it is vaporized resulting in a small amount of the diagnostic product being expelled into a flight tube. As the diagnostic product is charged when an electrical field pulse is subsequently applied to the tube they are launched down the flight tube towards a detector. The time between application of the electrical field pulse and collision of the diagnostic product with the detector is referred to as the time of flight. This is a very precise measure of the product's molecular weight, as a molecule's mass correlates directly with time of flight with smaller molecules flying faster than larger molecules. The entire assay is completed in less than one thousandth of a second, enabling samples to be analyzed in a total of 3-5 second including repetitive data collection. The SpectroTYPER software then calculates, records, compares and reports the genotypes at the rate of three seconds per sample.

7. Detection of Variant MCFD2 Proteins

In other embodiments, variant (*e.g.*, truncated) MCFD2 polypeptides are detected (*e.g.*, including, but not limited to, those described in SEQ ID NOs: 6, 8, 10, 12, 14, 16, and 18). Any suitable method may be used to detect truncated or mutant MCFD2 polypeptides including, but not limited to, those described below.

a) Cell Free Translation

For example, in some embodiments, cell-free translation methods from Amberg, Inc. (Boston, MA) are utilized. Amberg, Inc. has developed a method for the labeling, detection, quantitation, analysis and isolation of nascent proteins produced in a cell-free or cellular translation system without the use of radioactive amino acids or other radioactive labels. Markers are aminoacylated to tRNA molecules. Potential markers include native amino acids, non-native amino acids, amino acid analogs or derivatives, or chemical moieties. These markers are introduced into nascent proteins from the resulting misaminoacylated tRNAs during the translation process.

One application of Amberg's protein labeling technology is the gel free truncation test (GFTT) assay (See *e.g.*, U.S. Patent 6,303,337, herein incorporated by reference). In some embodiments, this assay is used to screen for truncation mutations in a TSC1 or TSC2 protein. In the GFTT assay, a marker (*e.g.*, a fluorophore) is introduced to the nascent protein during translation near the N-terminus of the protein. A second and different marker (*e.g.*, a fluorophore with a different emission wavelength) is introduced to the nascent protein near the C-terminus of the protein. The protein is then separated from the translation system and the signal from the markers is measured. A comparison of the measurements from the N and C terminal signals provides information on the fraction of the molecules with C-terminal truncation (*i.e.*, if the normalized signal from the C-terminal marker is 50% of the signal from the N-terminal marker, 50% of the molecules have a C-terminal truncation).

b) Antibody Binding

In still further embodiments of the present invention, antibodies (See below for antibody production) are used to determine if an individual contains an allele encoding a variant MCFD2 gene. In preferred embodiments, antibodies are utilized that discriminate between variant (*i.e.*, truncated proteins); and wild-type proteins (SEQ ID NOs:2). In some particularly preferred embodiments, the antibodies are directed to the C-terminus of MCFD2. Proteins that are recognized by the N-terminal, but not the C-terminal antibody are truncated. In some embodiments, quantitative immunoassays are used to determine the ratios of C-terminal to N-terminal antibody binding. In other embodiments, antibodies that differentially bind to wild type or variant forms of MCFD2.

Antibody binding is detected by techniques known in the art (*e.g.*, radioimmunoassay, ELISA (enzyme-linked immunosorbant assay), "sandwich" immunoassays, immunoradiometric assays, gel diffusion precipitation reactions, immunodiffusion assays, *in situ* immunoassays (*e.g.*, using colloidal gold, enzyme or radioisotope labels, for example), Western blots, precipitation reactions, agglutination assays (*e.g.*, gel agglutination assays, hemagglutination assays, etc.), complement fixation assays, immunofluorescence assays, protein A assays, and immunoelectrophoresis assays, etc.

In one embodiment, antibody binding is detected by detecting a label on the primary antibody. In another embodiment, the primary antibody is detected by detecting binding of a secondary antibody or reagent to the primary antibody. In a further embodiment, the secondary antibody is labeled. Many methods are known in the art for detecting binding in an immunoassay and are within the scope of the present invention.

In some embodiments, an automated detection assay is utilized. Methods for the automation of immunoassays include those described in U.S. Patents 5,885,530, 4,981,785, 6,159,750, and 5,358,691, each of which is herein incorporated by reference. In some embodiments, the analysis and presentation of results is also automated. For example, in some embodiments, software that generates a prognosis based on the result of the immunoassay is utilized.

In other embodiments, the immunoassay described in U.S. Patents 5,599,677 and 5,672,480; each of which is herein incorporated by reference.

8. Kits for Analyzing Risk of F5F8D

The present invention also provides kits for determining whether an individual contains a wild-type or variant (*e.g.*, mutant or polymorphic) allele of MCFD2. In some embodiments, the kits are useful determining whether the subject is at risk of developing F5F8D. The diagnostic kits are produced in a variety of ways. In some embodiments, the kits contain at least one reagent for specifically detecting a mutant MCFD2 allele or protein. In preferred embodiments, the kits contain reagents for detecting a truncation in the *MCFD2* gene. In preferred embodiments, the reagent is a nucleic acid that hybridizes to nucleic acids containing the mutation and that does not bind to nucleic acids that do not contain the mutation. In other preferred embodiments, the reagents are primers for amplifying the region of DNA containing the mutation. In still other embodiments, the reagents are antibodies that preferentially bind either the wild-type or truncated MCFD2 proteins.

In some embodiments, the kit contains instructions for determining whether the subject is at risk for developing F5F8D. In preferred embodiments, the instructions specify that risk for developing F5F8D is determined by detecting the presence or

absence of a mutant MCFD2 allele in the subject, wherein subjects having an mutant (e.g., truncated) allele are at greater risk for F5F8D disease.

The presence or absence of a disease-associated mutation in a *MCFD2* gene can be used for therapeutic or other medical decisions. For example, couples with a family history of F5F8D may choose to conceive a child via *in vitro* fertilization and pre-implantation genetic screening. In this case, fertilized embryos are screened for mutant (e.g., disease associated) alleles of the MCFD2 gene and only embryos with wild type alleles are implanted in the uterus.

In other embodiments, in utero screening is performed on a developing fetus (e.g., amniocentesis or chorionic villi screening). In still other embodiments, genetic screening of newborn babies or very young children is performed. The early detection of a MCFD2 allele known to be associated with F5F8D allows for early intervention (e.g., genetic or pharmaceutical therapies).

In some embodiments, the kits include ancillary reagents such as buffering agents, nucleic acid stabilizing reagents, protein stabilizing reagents, and signal producing systems (e.g., fluorescence generating systems as FRET systems). The test kit may be packages in any suitable manner, typically with the elements in a single container or various containers as necessary along with a sheet of instructions for carrying out the test. In some embodiments, the kits also preferably include a positive control sample.

20

9. Bioinformatics

In some embodiments, the present invention provides methods of determining an individual's risk of developing F5F8D based on the presence of one or more variant alleles of MCFD2. In some embodiments, the analysis of variant data is processed by a computer using information stored on a computer (e.g., in a database). For example, in some embodiments, the present invention provides a bioinformatics research system comprising a plurality of computers running a multi-platform object oriented programming language (See e.g., U.S. Patent 6,125,383; herein incorporated by reference). In some embodiments, one of the computers stores genetics data (e.g., the risk of contacting F5F8D associated with a given polymorphism, as well as the sequences). In some embodiments, one of the computers stores application programs

30

(*e.g.*, for analyzing the results of detection assays). Results are then delivered to the user (*e.g.*, via one of the computers or via the Internet).

For example, in some embodiments, a computer-based analysis program is used to translate the raw data generated by the detection assay (*e.g.*, the presence, absence, or amount of a given MCFD2 allele or polypeptide) into data of predictive value for a clinician. The clinician can access the predictive data using any suitable means. Thus, in some preferred embodiments, the present invention provides the further benefit that the clinician, who is not likely to be trained in genetics or molecular biology, need not understand the raw data. The data is presented directly to the clinician in its most useful form. The clinician is then able to immediately utilize the information in order to optimize the care of the subject.

The present invention contemplates any method capable of receiving, processing, and transmitting the information to and from laboratories conducting the assays, information provides, medical personal, and subjects. For example, in some embodiments of the present invention, a sample (*e.g.*, a biopsy or a serum or urine sample) is obtained from a subject and submitted to a profiling service (*e.g.*, clinical lab at a medical facility, genomic profiling business, etc.), located in any part of the world (*e.g.*, in a country different than the country where the subject resides or where the information is ultimately used) to generate raw data. Where the sample comprises a tissue or other biological sample, the subject may visit a medical center to have the sample obtained and sent to the profiling center, or subjects may collect the sample themselves (*e.g.*, a urine sample) and directly send it to a profiling center. Where the sample comprises previously determined biological information, the information may be directly sent to the profiling service by the subject (*e.g.*, an information card containing the information may be scanned by a computer and the data transmitted to a computer of the profiling center using an electronic communication systems). Once received by the profiling service, the sample is processed and a profile is produced (*i.e.*, presence of wild type or mutant MCFD2 genes or polypeptides), specific for the diagnostic or prognostic information desired for the subject.

The profile data is then prepared in a format suitable for interpretation by a treating clinician. For example, rather than providing raw data, the prepared format may

represent a diagnosis or risk assessment (*e.g.*, likelihood of developing F5F8D or a diagnosis of MCFD2 polymorphism) for the subject, along with recommendations for particular treatment options. The data may be displayed to the clinician by any suitable method. For example, in some embodiments, the profiling service generates a report that
5 can be printed for the clinician (*e.g.*, at the point of care) or displayed to the clinician on a computer monitor.

In some embodiments, the information is first analyzed at the point of care or at a regional facility. The raw data is then sent to a central processing facility for further analysis and/or to convert the raw data to information useful for a clinician or patient.

10 The central processing facility provides the advantage of privacy (all data is stored in a central facility with uniform security protocols), speed, and uniformity of data analysis. The central processing facility can then control the fate of the data following treatment of the subject. For example, using an electronic communication system, the central facility can provide data to the clinician, the subject, or researchers.

15 In some embodiments, the subject is able to directly access the data using the electronic communication system. The subject may chose further intervention or counseling based on the results. In some embodiments, the data is used for research use. For example, the data may be used to further optimize the inclusion or elimination of markers as useful indicators of a particular condition or stage of disease.

20

IV. Generation of MCFD2 Antibodies

The present invention provides isolated antibodies or antibody fragments (*e.g.*, FAB fragments). Antibodies can be generated to allow for the detection of MCFD2 protein. The antibodies may be prepared using various immunogens. In one
25 embodiment, the immunogen is a human MCFD2 peptide to generate antibodies that recognize human MCFD2. Such antibodies include, but are not limited to polyclonal, monoclonal, chimeric, single chain, Fab fragments, Fab expression libraries, or recombinant (*e.g.*, chimeric, humanized, etc.) antibodies, as long as it can recognize the protein. Antibodies can be produced by using a protein of the present invention as the
30 antigen according to a conventional antibody or antiserum preparation process.

Various procedures known in the art may be used for the production of polyclonal antibodies directed against MCFD2. For the production of antibody, various host animals can be immunized by injection with the peptide corresponding to the MCFD2 epitope including but not limited to rabbits, mice, rats, sheep, goats, etc. In a preferred embodiment, the peptide is conjugated to an immunogenic carrier (*e.g.*, diphtheria toxoid, bovine serum albumin (BSA), or keyhole limpet hemocyanin (KLH)). Various adjuvants may be used to increase the immunological response, depending on the host species, including but not limited to Freund's (complete and incomplete), mineral gels (*e.g.*, aluminum hydroxide), surface active substances (*e.g.*, lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanins, dinitrophenol, and potentially useful human adjuvants such as BCG (Bacille Calmette-Guerin) and *Corynebacterium parvum*).

For preparation of monoclonal antibodies directed toward MCFD2, it is contemplated that any technique that provides for the production of antibody molecules by continuous cell lines in culture will find use with the present invention (*See e.g.*, Harlow and Lane, *Antibodies: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY). These include but are not limited to the hybridoma technique originally developed by Köhler and Milstein (Köhler and Milstein, *Nature* 256:495-497 [1975]), as well as the trioma technique, the human B-cell hybridoma technique (*See e.g.*, Kozbor *et al.*, *Immunol. Tod.*, 4:72 [1983]), and the EBV-hybridoma technique to produce human monoclonal antibodies (Cole *et al.*, in *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, Inc., pp. 77-96 [1985]).

In an additional embodiment of the invention, monoclonal antibodies are produced in germ-free animals utilizing technology such as that described in PCT/US90/02545). Furthermore, it is contemplated that human antibodies will be generated by human hybridomas (Cote *et al.*, *Proc. Natl. Acad. Sci. USA* 80:2026-2030 [1983]) or by transforming human B cells with EBV virus *in vitro* (Cole *et al.*, in *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, pp. 77-96 [1985]).

In addition, it is contemplated that techniques described for the production of single chain antibodies (U.S. Patent 4,946,778; herein incorporated by reference) will find use in producing MCFD2 specific single chain antibodies. An additional

embodiment of the invention utilizes the techniques described for the construction of Fab expression libraries (Huse *et al.*, Science 246:1275-1281 [1989]) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity for MCFD2.

In other embodiments, the present invention contemplated recombinant antibodies
5 or fragments thereof to the proteins of the present invention. Recombinant antibodies include, but are not limited to, humanized and chimeric antibodies. Methods for generating recombinant antibodies are known in the art (*See e.g.*, U.S. Patents 6,180,370 and 6,277,969 and "Monoclonal Antibodies" H. Zola, BIOS Scientific Publishers Limited 2000. Springer-Verlay New York, Inc., New York; each of which is herein incorporated
10 by reference).

It is contemplated that any technique suitable for producing antibody fragments will find use in generating antibody fragments that contain the idiotype (antigen binding region) of the antibody molecule. For example, such fragments include but are not limited to: F(ab')₂ fragment that can be produced by pepsin digestion of the antibody
15 molecule; Fab' fragments that can be generated by reducing the disulfide bridges of the F(ab')₂ fragment, and Fab fragments that can be generated by treating the antibody molecule with papain and a reducing agent.

In the production of antibodies, it is contemplated that screening for the desired antibody will be accomplished by techniques known in the art (*e.g.*, radioimmunoassay,
20 ELISA (enzyme-linked immunosorbant assay), "sandwich" immunoassays, immunoradiometric assays, gel diffusion precipitation reactions, immunodiffusion assays, *in situ* immunoassays (*e.g.*, using colloidal gold, enzyme or radioisotope labels, for example), Western blots, precipitation reactions, agglutination assays (*e.g.*, gel agglutination assays, hemagglutination assays, etc.), complement fixation assays,
25 immunofluorescence assays, protein A assays, and immunoelectrophoresis assays, etc.

In one embodiment, antibody binding is detected by detecting a label on the primary antibody. In another embodiment, the primary antibody is detected by detecting binding of a secondary antibody or reagent to the primary antibody. In a further embodiment, the secondary antibody is labeled. Many means are known in the art for
30 detecting binding in an immunoassay and are within the scope of the present invention. As is well known in the art, the immunogenic peptide should be provided free of the

carrier molecule used in any immunization protocol. For example, if the peptide was conjugated to KLH, it may be conjugated to BSA, or used directly, in a screening assay.)

The foregoing antibodies can be used in methods known in the art relating to the localization and structure of MCFD2 (*e.g.*, for Western blotting), measuring levels thereof in appropriate biological samples, etc. The antibodies can be used to detect MCFD2 in a biological sample from an individual. The biological sample can be a biological fluid, such as, but not limited to, blood, serum, plasma, interstitial fluid, urine, cerebrospinal fluid, and the like, containing cells.

The biological samples can then be tested directly for the presence of human MCFD2 using an appropriate strategy (*e.g.*, ELISA or radioimmunoassay) and format (*e.g.*, microwells, dipstick (*e.g.*, as described in International Patent Publication WO 93/03367), etc. Alternatively, proteins in the sample can be size separated (*e.g.*, by polyacrylamide gel electrophoresis (PAGE), in the presence or not of sodium dodecyl sulfate (SDS), and the presence of MCFD2 detected by immunoblotting (Western blotting). Immunoblotting techniques are generally more effective with antibodies generated against a peptide corresponding to an epitope of a protein, and hence, are particularly suited to the present invention.

Another method uses antibodies as agents to alter signal transduction. Specific antibodies that bind to the binding domains of MCFD2 or other proteins involved in intracellular signaling can be used to inhibit the interaction between the various proteins and their interaction with other ligands. Antibodies that bind to the complex can also be used therapeutically to inhibit interactions of the protein complex in the signal transduction pathways leading to the various physiological and cellular effects of MCFD2. Such antibodies can also be used diagnostically to measure abnormal expression of MCFD2, or the aberrant formation of protein complexes, which may be indicative of a disease state.

V. Gene Therapy Using *MCFD2*

The present invention also provides methods and compositions suitable for gene therapy to alter *MCFD2* expression, production, or function. As described above, the present invention provides human *MCFD2* genes and provides methods of obtaining
5 *MCFD2* genes from other species. Thus, the methods described below are generally applicable across many species. In some embodiments, it is contemplated that the gene therapy is performed by providing a subject with a wild-type allele of *MCFD2* (*i.e.*, an allele that does not cause a F5F8D disease (*e.g.*, free of disease causing polymorphisms or mutations)). Subjects in need of such therapy are identified by the methods described
10 above.

Viral vectors commonly used for *in vivo* or *ex vivo* targeting and therapy procedures are DNA-based vectors and retroviral vectors. Methods for constructing and using viral vectors are known in the art (*See e.g.*, Miller and Rosman, BioTech., 7:980-990 [1992]). Preferably, the viral vectors are replication defective, that is, they are
15 unable to replicate autonomously in the target cell. In general, the genome of the replication defective viral vectors that are used within the scope of the present invention lack at least one region that is necessary for the replication of the virus in the infected cell. These regions can either be eliminated (in whole or in part), or be rendered non-functional by any technique known to a person skilled in the art. These techniques
20 include the total removal, substitution (by other sequences, in particular by the inserted nucleic acid), partial deletion or addition of one or more bases to an essential (for replication) region. Such techniques may be performed *in vitro* (*i.e.*, on the isolated DNA) or *in situ*, using the techniques of genetic manipulation or by treatment with mutagenic agents.

25 Preferably, the replication defective virus retains the sequences of its genome that are necessary for encapsidating the viral particles. DNA viral vectors include an attenuated or defective DNA viruses, including, but not limited to, herpes simplex virus (HSV), papillomavirus, Epstein Barr virus (EBV), adenovirus, adeno-associated virus (AAV), and the like. Defective viruses, that entirely or almost entirely lack viral genes,
30 are preferred, as defective virus is not infective after introduction into a cell. Use of defective viral vectors allows for administration to cells in a specific, localized area,

without concern that the vector can infect other cells. Thus, a specific tissue can be specifically targeted. Examples of particular vectors include, but are not limited to, a defective herpes virus 1 (HSV1) vector (Kaplitt *et al.*, Mol. Cell. Neurosci., 2:320-330 [1991]), defective herpes virus vector lacking a glycoprotein L gene (*See e.g.*, Patent Publication RD 371005 A), or other defective herpes virus vectors (*See e.g.*, WO 94/21807; and WO 92/05263); an attenuated adenovirus vector, such as the vector described by Stratford-Perricaudet *et al.* (J. Clin. Invest., 90:626-630 [1992]; *See also*, La Salle *et al.*, Science 259:988-990 [1993]); and a defective adeno-associated virus vector (Samulski *et al.*, J. Virol., 61:3096-3101 [1987]; Samulski *et al.*, J. Virol., 63:3822-3828 [1989]; and Lebkowski *et al.*, Mol. Cell. Biol., 8:3988-3996 [1988]).

Preferably, for *in vivo* administration, an appropriate immunosuppressive treatment is employed in conjunction with the viral vector (*e.g.*, adenovirus vector), to avoid immuno-deactivation of the viral vector and transfected cells. For example, immunosuppressive cytokines, such as interleukin-12 (IL-12), interferon-gamma (IFN- γ), or anti-CD4 antibody, can be administered to block humoral or cellular immune responses to the viral vectors. In addition, it is advantageous to employ a viral vector that is engineered to express a minimal number of antigens.

In a preferred embodiment, the vector is an adenovirus vector. Adenoviruses are eukaryotic DNA viruses that can be modified to efficiently deliver a nucleic acid of the invention to a variety of cell types. Various serotypes of adenovirus exist. Of these serotypes, preference is given, within the scope of the present invention, to type 2 or type 5 human adenoviruses (Ad 2 or Ad 5), or adenoviruses of animal origin (*See e.g.*, WO 94/26914). Those adenoviruses of animal origin that can be used within the scope of the present invention include adenoviruses of canine, bovine, murine (*e.g.*, Mav1, Beard *et al.*, Virol., 75-81 [1990]), ovine, porcine, avian, and simian (*e.g.*, SAV) origin. Preferably, the adenovirus of animal origin is a canine adenovirus, more preferably a CAV2 adenovirus (*e.g.* Manhattan or A26/61 strain (ATCC VR-800)).

Preferably, the replication defective adenoviral vectors of the invention comprise the ITRs, an encapsidation sequence and the nucleic acid of interest. Still more preferably, at least the E1 region of the adenoviral vector is non-functional. The deletion in the E1 region preferably extends from nucleotides 455 to 3329 in the sequence of the

Ad5 adenovirus (*PvuII-BglIII* fragment) or 382 to 3446 (*HinfII-Sau3A* fragment). Other regions may also be modified, in particular the E3 region (*e.g.*, WO 95/02697), the E2 region (*e.g.*, WO 94/28938), the E4 region (*e.g.*, WO 94/28152, WO 94/12649 and WO 95/02697), or in any of the late genes L1-L5.

5 In a preferred embodiment, the adenoviral vector has a deletion in the E1 region (Ad 1.0). Examples of E1-deleted adenoviruses are disclosed in EP 185,573, the contents of which are incorporated herein by reference. In another preferred embodiment, the adenoviral vector has a deletion in the E1 and E4 regions (Ad 3.0). Examples of E1/E4-deleted adenoviruses are disclosed in WO 95/02697 and WO
10 96/22378. In still another preferred embodiment, the adenoviral vector has a deletion in the E1 region into which the E4 region and the nucleic acid sequence are inserted.

The replication defective recombinant adenoviruses according to the invention can be prepared by any technique known to the person skilled in the art (*See e.g.*, Levrero *et al.*, Gene 101:195 [1991]; EP 185 573; and Graham, EMBO J., 3:2917 [1984]). In
15 particular, they can be prepared by homologous recombination between an adenovirus and a plasmid that carries, *inter alia*, the DNA sequence of interest. The homologous recombination is accomplished following co-transfection of the adenovirus and plasmid into an appropriate cell line. The cell line that is employed should preferably (i) be transformable by the elements to be used, and (ii) contain the sequences that are able to
20 complement the part of the genome of the replication defective adenovirus, preferably in integrated form in order to avoid the risks of recombination. Examples of cell lines that may be used are the human embryonic kidney cell line 293 (Graham *et al.*, J. Gen. Virol., 36:59 [1977]), which contains the left-hand portion of the genome of an Ad5 adenovirus (12%) integrated into its genome, and cell lines that are able to complement the E1 and
25 E4 functions, as described in applications WO 94/26914 and WO 95/02697.

Recombinant adenoviruses are recovered and purified using standard molecular biological techniques that are well known to one of ordinary skill in the art.

The adeno-associated viruses (AAV) are DNA viruses of relatively small size that can integrate, in a stable and site-specific manner, into the genome of the cells that they
30 infect. They are able to infect a wide spectrum of cells without inducing any effects on cellular growth, morphology or differentiation, and they do not appear to be involved in

human pathologies. The AAV genome has been cloned, sequenced and characterized. It encompasses approximately 4700 bases and contains an inverted terminal repeat (ITR) region of approximately 145 bases at each end, which serves as an origin of replication for the virus. The remainder of the genome is divided into two essential regions that

5 carry the encapsidation functions: the left-hand part of the genome, that contains the *rep* gene involved in viral replication and expression of the viral genes; and the right-hand part of the genome, that contains the *cap* gene encoding the capsid proteins of the virus.

The use of vectors derived from the AAVs for transferring genes *in vitro* and *in vivo* has been described (*See e.g.*, WO 91/18088; WO 93/09239; US Pat. No. 4,797,368; US Pat. No., 5,139,941; and EP 488 528, all of which are herein incorporated by

10 reference). These publications describe various AAV-derived constructs in which the *rep* and/or *cap* genes are deleted and replaced by a gene of interest, and the use of these constructs for transferring the gene of interest *in vitro* (into cultured cells) or *in vivo* (directly into an organism). The replication defective recombinant AAVs according to

15 the invention can be prepared by co-transfecting a plasmid containing the nucleic acid sequence of interest flanked by two AAV inverted terminal repeat (ITR) regions, and a plasmid carrying the AAV encapsidation genes (*rep* and *cap* genes), into a cell line that is infected with a human helper virus (for example an adenovirus). The AAV recombinants that are produced are then purified by standard techniques.

20 In another embodiment, the gene can be introduced in a retroviral vector (*e.g.*, as described in U.S. Pat. Nos. 5,399,346, 4,650,764, 4,980,289 and 5,124,263; all of which are herein incorporated by reference; Mann *et al.*, Cell 33:153 [1983]; Markowitz *et al.*, J. Virol., 62:1120 [1988]; PCT/US95/14575; EP 453242; EP178220; Bernstein *et al.* Genet. Eng., 7:235 [1985]; McCormick, BioTechnol., 3:689 [1985]; WO 95/07358; and

25 Kuo *et al.*, Blood 82:845 [1993]). The retroviruses are integrating viruses that infect dividing cells. The retrovirus genome includes two LTRs, an encapsidation sequence and three coding regions (*gag*, *pol* and *env*). In recombinant retroviral vectors, the *gag*, *pol* and *env* genes are generally deleted, in whole or in part, and replaced with a heterologous nucleic acid sequence of interest. These vectors can be constructed from different types

30 of retrovirus, such as, HIV, MoMuLV ("murine Moloney leukemia virus" MSV ("murine Moloney sarcoma virus"), HaSV ("Harvey sarcoma virus"); SNV ("spleen necrosis

virus"); RSV ("Rous sarcoma virus") and Friend virus. Defective retroviral vectors are also disclosed in WO 95/02697.

In general, in order to construct recombinant retroviruses containing a nucleic acid sequence, a plasmid is constructed that contains the LTRs, the encapsidation
5 sequence and the coding sequence. This construct is used to transfect a packaging cell line, which cell line is able to supply in trans the retroviral functions that are deficient in the plasmid. In general, the packaging cell lines are thus able to express the *gag*, *pol* and *env* genes. Such packaging cell lines have been described in the prior art, in particular the cell line PA317 (US Pat. No. 4,861,719, herein incorporated by reference), the
10 PsiCRIP cell line (*See*, WO90/02806), and the GP+envAm-12 cell line (*See*, WO89/07150). In addition, the recombinant retroviral vectors can contain modifications within the LTRs for suppressing transcriptional activity as well as extensive encapsidation sequences that may include a part of the *gag* gene (Bender *et al.*, J. Virol., 61:1639 [1987]). Recombinant retroviral vectors are purified by standard techniques
15 known to those having ordinary skill in the art.

Alternatively, the vector can be introduced *in vivo* by lipofection. For the past decade, there has been increasing use of liposomes for encapsulation and transfection of nucleic acids *in vitro*. Synthetic cationic lipids designed to limit the difficulties and dangers encountered with liposome mediated transfection can be used to prepare
20 liposomes for *in vivo* transfection of a gene encoding a marker (Felgner *et al.*, Proc. Natl. Acad. Sci. USA 84:7413-7417 [1987]; *See also*, Mackey, *et al.*, Proc. Natl. Acad. Sci. USA 85:8027-8031 [1988]; Ulmer *et al.*, Science 259:1745-1748 [1993]). The use of cationic lipids may promote encapsulation of negatively charged nucleic acids, and also promote fusion with negatively charged cell membranes (Felgner and Ringold, Science
25 337:387-388 [1989]). Particularly useful lipid compounds and compositions for transfer of nucleic acids are described in WO95/18863 and WO96/17823, and in U.S. Pat. No. 5,459,127, herein incorporated by reference.

Other molecules are also useful for facilitating transfection of a nucleic acid *in vivo*, such as a cationic oligopeptide (*e.g.*, WO95/21931), peptides derived from DNA
30 binding proteins (*e.g.*, WO96/25508), or a cationic polymer (*e.g.*, WO95/21931).

It is also possible to introduce the vector *in vivo* as a naked DNA

plasmid. Methods for formulating and administering naked DNA to mammalian muscle tissue are disclosed in U.S. Pat. Nos. 5,580,859 and 5,589,466, both of which are herein incorporated by reference.

DNA vectors for gene therapy can be introduced into the desired host cells by methods known in the art, including but not limited to transfection, electroporation, microinjection, transduction, cell fusion, DEAE dextran, calcium phosphate precipitation, use of a gene gun, or use of a DNA vector transporter (*See e.g.*, Wu *et al.*, J. Biol. Chem., 267:963 [1992]; Wu and Wu, J. Biol. Chem., 263:14621 [1988]; and Williams *et al.*, Proc. Natl. Acad. Sci. USA 88:2726 [1991]). Receptor-mediated DNA delivery approaches can also be used (Curiel *et al.*, Hum. Gene Ther., 3:147 [1992]; and Wu and Wu, J. Biol. Chem., 262:4429 [1987]).

VI. Transgenic Animals Expressing Exogenous *MCFD2* Genes and Homologs, Mutants, and Variants Thereof

The present invention contemplates the generation of transgenic animals comprising an exogenous *MCFD2* gene or homologs, mutants, or variants thereof. In preferred embodiments, the transgenic animal displays an altered phenotype as compared to wild-type animals. In some embodiments, the altered phenotype is the overexpression of mRNA for a *MCFD2* gene as compared to wild-type levels of *MCFD2* expression. In other embodiments, the altered phenotype is the decreased expression of mRNA for an endogenous *MCFD2* gene as compared to wild-type levels of endogenous *MCFD2* expression. In some preferred embodiments, the transgenic animals comprise mutant (*e.g.*, truncated) alleles of *MCFD2*, in the presence or absence of the corresponding wild-type allele. Methods for analyzing the presence or absence of such phenotypes include Northern blotting, mRNA protection assays, and RT-PCR. In other embodiments, the transgenic mice have a knock out mutation of the *MCFD2* gene. In preferred embodiments, the transgenic animals display a F5F8D phenotype.

Such animals find use in research applications (*e.g.*, identifying signaling pathways that *MCFD2* is involved in), as well as drug screening applications (*e.g.*, to screen for drugs that prevents F5F8D). For example, in some embodiments, test compounds (*e.g.*, a drug that is suspected of being useful to treat F5F8D) and control

compounds (*e.g.*, a placebo) are administered to the transgenic animals and the control animals and the effects evaluated. The effects of the test and control compounds on disease symptoms are then assessed.

The transgenic animals can be generated via a variety of methods. In some
5 embodiments, embryonal cells at various developmental stages are used to introduce
transgenes for the production of transgenic animals. Different methods are used
depending on the stage of development of the embryonal cell. The zygote is the best
target for micro-injection. In the mouse, the male pronucleus reaches the size of
approximately 20 micrometers in diameter, which allows reproducible injection of 1-2
10 picoliters (pl) of DNA solution. The use of zygotes as a target for gene transfer has a
major advantage in that in most cases the injected DNA will be incorporated into the host
genome before the first cleavage (Brinster *et al.*, Proc. Natl. Acad. Sci. USA 82:4438-
4442 [1985]). As a consequence, all cells of the transgenic non-human animal will carry
the incorporated transgene. This will in general also be reflected in the efficient
15 transmission of the transgene to offspring of the founder since 50% of the germ cells will
harbor the transgene. U.S. Patent No. 4,873,191 describes a method for the micro-
injection of zygotes; the disclosure of this patent is incorporated herein in its entirety.

In other embodiments, retroviral infection is used to introduce transgenes into a
non-human animal. In some embodiments, the retroviral vector is utilized to transfect
20 oocytes by injecting the retroviral vector into the perivitelline space of the oocyte (U.S.
Pat. No. 6,080,912, incorporated herein by reference). In other embodiments, the
developing non-human embryo can be cultured *in vitro* to the blastocyst stage. During
this time, the blastomeres can be targets for retroviral infection (Janenich, Proc. Natl.
Acad. Sci. USA 73:1260 [1976]). Efficient infection of the blastomeres is obtained by
25 enzymatic treatment to remove the zona pellucida (Hogan *et al.*, in *Manipulating the
Mouse Embryo*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. [1986]).
The viral vector system used to introduce the transgene is typically a replication-defective
retrovirus carrying the transgene (Jahner *et al.*, Proc. Natl. Acad. Sci. USA 82:6927
[1985]). Transfection is easily and efficiently obtained by culturing the blastomeres on a
30 monolayer of virus-producing cells (Van der Putten, *supra*; Stewart, *et al.*, EMBO J.,
6:383 [1987]). Alternatively, infection can be performed at a later stage. Virus or virus-

producing cells can be injected into the blastocoele (Jahner *et al.*, Nature 298:623 [1982]). Most of the founders will be mosaic for the transgene since incorporation occurs only in a subset of cells that form the transgenic animal. Further, the founder may contain various retroviral insertions of the transgene at different positions in the genome that generally will segregate in the offspring. In addition, it is also possible to introduce transgenes into the germline, albeit with low efficiency, by intrauterine retroviral infection of the midgestation embryo (Jahner *et al.*, *supra* [1982]). Additional means of using retroviruses or retroviral vectors to create transgenic animals known to the art involves the micro-injection of retroviral particles or mitomycin C-treated cells producing retrovirus into the perivitelline space of fertilized eggs or early embryos (PCT International Application WO 90/08832 [1990], and Haskell and Bowen, Mol. Reprod. Dev., 40:386 [1995]).

In other embodiments, the transgene is introduced into embryonic stem cells and the transfected stem cells are utilized to form an embryo. ES cells are obtained by culturing pre-implantation embryos *in vitro* under appropriate conditions (Evans *et al.*, Nature 292:154 [1981]; Bradley *et al.*, Nature 309:255 [1984]; Gossler *et al.*, Proc. Acad. Sci. USA 83:9065 [1986]; and Robertson *et al.*, Nature 322:445 [1986]). Transgenes can be efficiently introduced into the ES cells by DNA transfection by a variety of methods known to the art including calcium phosphate co-precipitation, protoplast or spheroplast fusion, lipofection and DEAE-dextran-mediated transfection. Transgenes may also be introduced into ES cells by retrovirus-mediated transduction or by micro-injection. Such transfected ES cells can thereafter colonize an embryo following their introduction into the blastocoele of a blastocyst-stage embryo and contribute to the germ line of the resulting chimeric animal (for review, *See*, Jaenisch, Science 240:1468 [1988]). Prior to the introduction of transfected ES cells into the blastocoele, the transfected ES cells may be subjected to various selection protocols to enrich for ES cells which have integrated the transgene assuming that the transgene provides a means for such selection.

Alternatively, the polymerase chain reaction may be used to screen for ES cells that have integrated the transgene. This technique obviates the need for growth of the transfected ES cells under appropriate selective conditions prior to transfer into the blastocoele.

In still other embodiments, homologous recombination is utilized to knock-out gene function or create deletion mutants. Methods for homologous recombination are described in U.S. Pat. No. 5,614,396, incorporated herein by reference.

5 **VIII. Drug Screening Using MCFD2**

As described herein, it is contemplated that MCFD2 and ERGIC-53 interact within a novel shared pathogenic pathway. Accordingly, in some embodiments, the isolated nucleic acid and protein sequences of *MCFD2* are used in drug screening applications for compounds that alter (e.g., enhance or inhibit) transport along the
10 pathway. In other embodiments, cells or tissues containing mutant MCFD2 sequences are tested with compounds (e.g., drugs, expression vectors, etc.) to identify factors that compensate for mutant MCFD2.

In some embodiments, compounds (e.g., drugs, antisense oligonucleotide, siRNAs, etc.) are identified that inhibit MCFD2 biological activity by targeting MCFD2
15 and/or one or more other proteins in a MCFD2 biological pathway (e.g., LMAN1). It is contemplated that such compounds find use as anticoagulants for treatment of diseases and conditions (e.g., thrombosis).

A. Identification of Binding Partners

20 In some embodiments, binding partners of MCFD2 amino acids are identified. In some embodiments, the MCFD2 nucleic acid sequence (e.g., SEQ ID NOS: 1, 5, 7, 9, 11, 13, 15, and 17) or fragments thereof are used in yeast two-hybrid screening assays. For example, in some embodiments, the nucleic acid sequences are subcloned into *pGPT9* (Clontech, La Jolla, CA) to be used as a bait in a yeast-2-hybrid screen for protein-protein
25 interaction of a human liver or megakaryocyte cDNA library (Fields and Song *Nature* 340:245 -246, 1989; herein incorporated by reference). In other embodiments, phage display is used to identify binding partners (Parmley and Smith *Gene* 73 : 305-318, [1988]; herein incorporated by reference).

30 **B. Drug Screening**

The present invention provides methods and compositions for using MCFD2 as a target for screening drugs that can alter, for example, interaction between MCFD2 and MCFD2 binding partners (*e.g.*, those identified using the above methods)

In one screening method, the two-hybrid system is used to screen for compounds (*e.g.*, drug) capable of altering (*e.g.*, inhibiting) MCFD2 function(s) (*e.g.*, interaction with a binding partner) *in vitro* or *in vivo*. In one embodiment, a GAL4 binding site, linked to a reporter gene such as *lacZ*, is contacted in the presence and absence of a candidate compound with a GAL4 binding domain linked to a MCFD2 fragment and a GAL4 transactivation domain II linked to a binding partner fragment. Expression of the reporter gene is monitored and a decrease in the expression is an indication that the candidate compound inhibits the interaction of MCFD2 with the binding partner. Alternately, the effect of candidate compounds on the interaction of MCFD2 with other proteins (*e.g.*, proteins known to interact directly or indirectly with the binding partner) can be tested in a similar manner.

In another screening method, candidate compounds are evaluated for their ability to alter MCFD2 transport by contacting MCFD2, binding partners, binding partner-associated proteins, or fragments thereof, with the candidate compound and determining binding of the candidate compound to the peptide. The protein or protein fragments is/are immobilized using methods known in the art such as binding a GST- MCFD2 fusion protein to a polymeric bead containing glutathione. A chimeric gene encoding a GST fusion protein is constructed by fusing DNA encoding the polypeptide or polypeptide fragment of interest to the DNA encoding the carboxyl terminus of GST (*See e.g.*, Smith *et al.*, Gene 67:31 [1988]). The fusion construct is then transformed into a suitable expression system (*e.g.*, *E. coli* XA90) in which the expression of the GST fusion protein can be induced with isopropyl- β -D-thiogalactopyranoside (IPTG). Induction with IPTG should yield the fusion protein as a major constituent of soluble, cellular proteins. The fusion proteins can be purified by methods known to those skilled in the art, including purification by glutathione affinity chromatography. Binding of the candidate compound to the proteins or protein fragments is correlated with the ability of the compound to disrupt the ERGIC pathway and thus regulate MCFD2 physiological effects (*e.g.*, F5F8D).

In another screening method, one of the components of the MCFD2/binding partner signaling system, is immobilized. Polypeptides can be immobilized using methods known in the art, such as adsorption onto a plastic microtiter plate or specific binding of a GST-fusion protein to a polymeric bead containing glutathione. For example, GST- MCFD2 is bound to glutathione-Sepharose beads. The immobilized peptide is then contacted with another peptide with which it is capable of binding in the presence and absence of a candidate compound. Unbound peptide is then removed and the complex solubilized and analyzed to determine the amount of bound labeled peptide. A decrease in binding is an indication that the candidate compound inhibits the interaction of MCFD2 with the other peptide. A variation of this method allows for the screening of compounds that are capable of disrupting a previously-formed protein/protein complex. For example, in some embodiments a complex comprising MCFD2 or a MCFD2 fragment bound to another peptide is immobilized as described above and contacted with a candidate compound. The dissolution of the complex by the candidate compound correlates with the ability of the compound to disrupt or inhibit the interaction between MCFD2 and the other peptide.

Another technique for drug screening provides high throughput screening for compounds having suitable binding affinity to MCFD2 peptides and is described in detail in WO 84/03564, incorporated herein by reference. Briefly, large numbers of different small peptide test compounds are synthesized on a solid substrate, such as plastic pins or some other surface. The peptide test compounds are then reacted with MCFD2 peptides and washed. Bound MCFD2 peptides are then detected by methods well known in the art.

Another technique uses MCFD2 antibodies, generated as discussed above. Such antibodies capable of specifically binding to MCFD2 peptides compete with a test compound for binding to MCFD2. In this manner, the antibodies can be used to detect the presence of any peptide that shares one or more antigenic determinants of the MCFD2 peptide.

The present invention contemplates many other means of screening compounds. The examples provided above are presented merely to illustrate a range of techniques

available. One of ordinary skill in the art will appreciate that many other screening methods can be used.

In particular, the present invention contemplates the use of cell lines transfected with MCFD2 and variants thereof for screening compounds for activity, and in particular to high throughput screening of compounds from combinatorial libraries (*e.g.*, libraries containing greater than 10^4 compounds). The cell lines of the present invention can be used in a variety of screening methods. In some embodiments, the cells can be used in second messenger assays that monitor signal transduction following activation of cell-surface receptors. In other embodiments, the cells can be used in reporter gene assays that monitor cellular responses at the transcription/translation level. In still further embodiments, the cells can be used in cell proliferation assays to monitor the overall growth/no growth response of cells to external stimuli.

In second messenger assays, the host cells are preferably transfected as described above with vectors encoding MCFD2 or variants or mutants thereof. The host cells are then treated with a compound or plurality of compounds (*e.g.*, from a combinatorial library) and assayed for the presence or absence of a response. It is contemplated that at least some of the compounds in the combinatorial library can serve as agonists, antagonists, activators, or inhibitors of the protein or proteins encoded by the vectors. It is also contemplated that at least some of the compounds in the combinatorial library can serve as agonists, antagonists, activators, or inhibitors of protein acting upstream or downstream of the protein encoded by the vector in a signal transduction pathway.

In some embodiments, the second messenger assays measure fluorescent signals from reporter molecules that respond to intracellular changes (*e.g.*, Ca^{2+} concentration, membrane potential, pH, IP_3 , cAMP, arachidonic acid release) due to stimulation of membrane receptors and ion channels (*e.g.*, ligand gated ion channels; *see Denyer et al.*, Drug Discov. Today 3:323 [1998]; and Gonzales *et al.*, Drug. Discov. Today 4:431-39 [1999]). Examples of reporter molecules include, but are not limited to, FRET (fluorescence resonance energy transfer) systems (*e.g.*, Cuo-lipids and oxonols, EDAN/DABCYL), calcium sensitive indicators (*e.g.*, Fluo-3, FURA 2, INDO 1, and FLUO3/AM, BAPTA AM), chloride-sensitive indicators (*e.g.*, SPQ, SPA), potassium-

sensitive indicators (*e.g.*, PBFI), sodium-sensitive indicators (*e.g.*, SBFI), and pH sensitive indicators (*e.g.*, BCECF).

In general, the host cells are loaded with the indicator prior to exposure to the compound. Responses of the host cells to treatment with the compounds can be detected
5 by methods known in the art, including, but not limited to, fluorescence microscopy, confocal microscopy (*e.g.*, FCS systems), flow cytometry, microfluidic devices, FLIPR systems (*See, e.g.*, Schroeder and Neagle, J. Biomol. Screening 1:75 [1996]), and plate-reading systems. In some preferred embodiments, the response (*e.g.*, increase in fluorescent intensity) caused by compound of unknown activity is compared to the
10 response generated by a known agonist and expressed as a percentage of the maximal response of the known agonist. The maximum response caused by a known agonist is defined as a 100% response. Likewise, the maximal response recorded after addition of an agonist to a sample containing a known or test antagonist is detectably lower than the 100% response.

15 The cells are also useful in reporter gene assays. Reporter gene assays involve the use of host cells transfected with vectors encoding a nucleic acid comprising transcriptional control elements of a target gene (*i.e.*, a gene that controls the biological expression and function of a disease target) spliced to a coding sequence for a reporter gene. Therefore, activation of the target gene results in activation of the reporter gene
20 product. In some embodiments, the reporter gene construct comprises the 5' regulatory region (*e.g.*, promoters and/or enhancers) of a protein whose expression is controlled by MCFD2 in operable association with a reporter gene (See Example 4 and Inohara *et al.*, J. Biol. Chem. 275:27823 [2000] for a description of the luciferase reporter construct pBVix-Luc). Examples of reporter genes finding use in the present invention include,
25 but are not limited to, chloramphenicol transferase, alkaline phosphatase, firefly and bacterial luciferases, β -galactosidase, β -lactamase, and green fluorescent protein. The production of these proteins, with the exception of green fluorescent protein, is detected through the use of chemiluminescent, colorimetric, or bioluminescent products of specific substrates (*e.g.*, X-gal and luciferin). Comparisons between compounds of known and
30 unknown activities may be conducted as described above.

Specifically, the present invention provides screening methods for identifying modulators, *i.e.*, candidate or test compounds or agents (*e.g.*, proteins, peptides, peptidomimetics, peptoids, small molecules or other drugs) which bind to MCFD2 of the present invention, have an inhibitory (or stimulatory) effect on, for example, MCFD2

5 expression or MCFD2 activity, or have a stimulatory or inhibitory effect on, for example, the expression or activity of a MCFD2 substrate. Compounds thus identified can be used to modulate the activity of target gene products (*e.g.*, *MCFD2* genes) either directly or indirectly in a therapeutic protocol, to elaborate the biological function of the target gene product, or to identify compounds that disrupt normal target gene interactions.

10 Compounds which stimulate the activity of a variant MCFD2 or mimic the activity of a non-functional variant are particularly useful in the treatment of hematologic diseases (*e.g.*, F5F8D).

In one embodiment, the invention provides assays for screening candidate or test compounds that are substrates of a MCFD2 protein or polypeptide or a biologically active

15 portion thereof. In another embodiment, the invention provides assays for screening candidate or test compounds that bind to or modulate the activity of a MCFD2 protein or polypeptide or a biologically active portion thereof.

The test compounds of the present invention can be obtained using any of the numerous approaches in combinatorial library methods known in the art, including

20 biological libraries; peptoid libraries (libraries of molecules having the functionalities of peptides, but with a novel, non-peptide backbone, which are resistant to enzymatic degradation but which nevertheless remain bioactive; see, *e.g.*, Zuckermann *et al.*, *J. Med. Chem.* 37: 2678-85 [1994]); spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; the 'one-bead one-

25 compound' library method; and synthetic library methods using affinity chromatography selection. The biological library and peptoid library approaches are preferred for use with peptide libraries, while the other four approaches are applicable to peptide, non-peptide oligomer or small molecule libraries of compounds (Lam (1997) *Anticancer Drug Des.* 12:145).

30 Examples of methods for the synthesis of molecular libraries can be found in the art, for example in: DeWitt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 90:6909 [1993]; Erb *et*

al., Proc. Nad. Acad. Sci. USA 91:11422 [1994]; Zuckermann *et al.*, J. Med. Chem. 37:2678 [1994]; Cho *et al.*, Science 261:1303 [1993]; Carrell *et al.*, Angew. Chem. Int. Ed. Engl. 33:2059 [1994]; Carell *et al.*, Angew. Chem. Int. Ed. Engl. 33:2061 [1994]; and Gallop *et al.*, J. Med. Chem. 37:1233 [1994].

5 Libraries of compounds may be presented in solution (*e.g.*, Houghten, Biotechniques 13:412-421 [1992]), or on beads (Lam, Nature 354:82-84 [1991]), chips (Fodor, Nature 364:555-556 [1993]), bacteria or spores (U.S. Patent No. 5,223,409; herein incorporated by reference), plasmids (Cull *et al.*, Proc. Nad. Acad. Sci. USA 89:1865-1869 [1992]) or on phage (Scott and Smith, Science 249:386-390 [1990]; Devlin
10 Science 249:404-406 [1990]; Cwirla *et al.*, Proc. Natl. Acad. Sci. 87:6378-6382 [1990]; Felici, J. Mol. Biol. 222:301 [1991]).

In one embodiment, an assay is a cell-based assay in which a cell that expresses a MCFD2 protein or biologically active portion thereof is contacted with a test compound, and the ability of the test compound to modulate MCFD2's activity is determined.

15 Determining the ability of the test compound to modulate MCFD2 activity can be accomplished by monitoring, for example, changes in enzymatic activity. The cell, for example, can be of mammalian origin.

The ability of the test compound to modulate MCFD2 binding to a compound, *e.g.*, a MCFD2 substrate, can also be evaluated. This can be accomplished, for example,
20 by coupling the compound, *e.g.*, the substrate, with a radioisotope or enzymatic label such that binding of the compound, *e.g.*, the substrate, to a MCFD2 can be determined by detecting the labeled compound, *e.g.*, substrate, in a complex.

Alternatively, the MCFD2 is coupled with a radioisotope or enzymatic label to monitor the ability of a test compound to modulate MCFD2 binding to a MCFD2
25 substrate in a complex. For example, compounds (*e.g.*, substrates) can be labeled with ¹²⁵I, ³⁵S, ¹⁴C or ³H, either directly or indirectly, and the radioisotope detected by direct counting of radioemission or by scintillation counting. Alternatively, compounds can be enzymatically labeled with, for example, horseradish peroxidase, alkaline phosphatase, or luciferase, and the enzymatic label detected by determination of
30 conversion of an appropriate substrate to product.

The ability of a compound (*e.g.*, a MCFD2 substrate) to interact with a MCFD2 with or without the labeling of any of the interactants can be evaluated. For example, a microphysiometer can be used to detect the interaction of a compound with a MCFD2 without the labeling of either the compound or the MCFD2 (McConnell *et al.* Science 257:1906-1912 [1992]). As used herein, a "microphysiometer" (*e.g.*, Cytosensor) is an analytical instrument that measures the rate at which a cell acidifies its environment using a light-addressable potentiometric sensor (LAPS). Changes in this acidification rate can be used as an indicator of the interaction between a compound and MCFD2.

In yet another embodiment, a cell-free assay is provided in which a MCFD2 protein or biologically active portion thereof is contacted with a test compound and the ability of the test compound to bind to the MCFD2 protein or biologically active portion thereof is evaluated. Preferred biologically active portions of the MCFD2 proteins to be used in assays of the present invention include fragments that participate in interactions with substrates or other proteins, *e.g.*, fragments with high surface probability scores.

Cell-free assays involve preparing a reaction mixture of the target gene protein and the test compound under conditions and for a time sufficient to allow the two components to interact and bind, thus forming a complex that can be removed and/or detected.

The interaction between two molecules can also be detected, *e.g.*, using fluorescence energy transfer (FRET) (see, for example, Lakowicz *et al.*, U.S. Patent No. 5,631,169; Stavrianopoulos *et al.*, U.S. Patent No. 4,968,103; each of which is herein incorporated by reference). A fluorophore label is selected such that a first donor molecule's emitted fluorescent energy will be absorbed by a fluorescent label on a second, 'acceptor' molecule, which in turn is able to fluoresce due to the absorbed energy.

Alternately, the 'donor' protein molecule may simply utilize the natural fluorescent energy of tryptophan residues. Labels are chosen that emit different wavelengths of light, such that the 'acceptor' molecule label may be differentiated from that of the 'donor'. Since the efficiency of energy transfer between the labels is related to the distance separating the molecules, the spatial relationship between the molecules can be assessed. In a situation in which binding occurs between the molecules, the fluorescent emission of the 'acceptor' molecule label in the assay should be maximal. An FRET

binding event can be conveniently measured through standard fluorometric detection means well known in the art (*e.g.*, using a fluorimeter).

In another embodiment, determining the ability of the MCFD2 protein to bind to a target molecule can be accomplished using real-time Biomolecular Interaction Analysis (BIA) (*see, e.g.*, Sjolander and Urbaniczky, *Anal. Chem.* 63:2338-2345 [1991] and Szabo *et al.* *Curr. Opin. Struct. Biol.* 5:699-705 [1995]). "Surface plasmon resonance" or "BIA" detects biospecific interactions in real time, without labeling any of the interactants (*e.g.*, BIAcore). Changes in the mass at the binding surface (indicative of a binding event) result in alterations of the refractive index of light near the surface (the optical phenomenon of surface plasmon resonance (SPR)), resulting in a detectable signal that can be used as an indication of real-time reactions between biological molecules.

In one embodiment, the target gene product or the test substance is anchored onto a solid phase. The target gene product/test compound complexes anchored on the solid phase can be detected at the end of the reaction. Preferably, the target gene product can be anchored onto a solid surface, and the test compound, (which is not anchored), can be labeled, either directly or indirectly, with detectable labels discussed herein.

It may be desirable to immobilize MCFD2, an anti- MCFD2 antibody or its target molecule to facilitate separation of complexed from non-complexed forms of one or both of the proteins, as well as to accommodate automation of the assay. Binding of a test compound to a MCFD2 protein, or interaction of a MCFD2 protein with a target molecule in the presence and absence of a candidate compound, can be accomplished in any vessel suitable for containing the reactants. Examples of such vessels include microtiter plates, test tubes, and micro-centrifuge tubes. In one embodiment, a fusion protein can be provided which adds a domain that allows one or both of the proteins to be bound to a matrix. For example, glutathione-S-transferase- MCFD2 fusion proteins or glutathione-S-transferase/target fusion proteins can be adsorbed onto glutathione Sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione-derivatized microtiter plates, which are then combined with the test compound or the test compound and either the non-adsorbed target protein or MCFD2 protein, and the mixture incubated under conditions conducive for complex formation (*e.g.*, at physiological conditions for salt and pH). Following incubation, the beads or microtiter plate wells are washed to remove any

unbound components, the matrix immobilized in the case of beads, complex determined either directly or indirectly, for example, as described above.

Alternatively, the complexes can be dissociated from the matrix, and the level of MCFD2 binding or activity determined using standard techniques. Other techniques for immobilizing either MCFD2 protein or a target molecule on matrices include using conjugation of biotin and streptavidin. Biotinylated MCFD2 protein or target molecules can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques known in the art (*e.g.*, biotinylation kit, Pierce Chemicals, Rockford, IL), and immobilized in the wells of streptavidin-coated 96 well plates (Pierce Chemical).

In order to conduct the assay, the non-immobilized component is added to the coated surface containing the anchored component. After the reaction is complete, unreacted components are removed (*e.g.*, by washing) under conditions such that any complexes formed will remain immobilized on the solid surface. The detection of complexes anchored on the solid surface can be accomplished in a number of ways.

Where the previously non-immobilized component is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the previously non-immobilized component is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; *e.g.*, using a labeled antibody specific for the immobilized component (the antibody, in turn, can be directly labeled or indirectly labeled with, *e.g.*, a labeled anti-IgG antibody).

This assay is performed utilizing antibodies reactive with MCFD2 protein or target molecules but which do not interfere with binding of the MCFD2 protein to its target molecule. Such antibodies can be derivatized to the wells of the plate, and unbound target or MCFD2 protein trapped in the wells by antibody conjugation. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the MCFD2 protein or target molecule, as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the MCFD2 protein or target molecule.

Alternatively, cell free assays can be conducted in a liquid phase. In such an assay, the reaction products are separated from unreacted components, by any of a

number of standard techniques, including, but not limited to: differential centrifugation (see, for example, Rivas and Minton, Trends Biochem Sci 18:284-7 [1993]); chromatography (gel filtration chromatography, ion-exchange chromatography); electrophoresis (*see, e.g.*, Ausubel *et al.*, eds. Current Protocols in Molecular Biology 1999, J. Wiley: New York.); and immunoprecipitation (*see, for example*, Ausubel *et al.*, eds. Current Protocols in Molecular Biology 1999, J. Wiley: New York). Such resins and chromatographic techniques are known to one skilled in the art (*See e.g.*, Heegaard J. Mol. Recognit 11:141-8 [1998]; Hageand Tweed J. Chromatogr. Biomed. Sci. Appl 699:499-525 [1997]). Further, fluorescence energy transfer may also be conveniently utilized, as described herein, to detect binding without further purification of the complex from solution.

The assay can include contacting the MCFD2 protein or biologically active portion thereof with a known compound that binds the MCFD2 to form an assay mixture, contacting the assay mixture with a test compound, and determining the ability of the test compound to interact with a MCFD2 protein, wherein determining the ability of the test compound to interact with a MCFD2 protein includes determining the ability of the test compound to preferentially bind to MCFD2 or biologically active portion thereof, or to modulate the activity of a target molecule, as compared to the known compound.

To the extent that MCFD2 can, *in vivo*, interact with one or more cellular or extracellular macromolecules, such as proteins, inhibitors of such an interaction are useful. A homogeneous assay can be used can be used to identify inhibitors.

For example, a preformed complex of the target gene product and the interactive cellular or extracellular binding partner product is prepared such that either the target gene products or their binding partners are labeled, but the signal generated by the label is quenched due to complex formation (*see, e.g.*, U.S. Patent No. 4,109,496, herein incorporated by reference, that utilizes this approach for immunoassays). The addition of a test substance that competes with and displaces one of the species from the preformed complex will result in the generation of a signal above background. In this way, test substances that disrupt target gene product-binding partner interaction can be identified. Alternatively, MCFD2 protein can be used as a "bait protein" in a two-hybrid assay or three-hybrid assay (*see, e.g.*, U.S. Patent No. 5,283,317; Zervos *et al.*, Cell 72:223-232

[1993]; Madura *et al.*, J. Biol. Chem. 268:12046-12054 [1993]; Bartel *et al.*, Biotechniques 14:920-924 [1993]; Iwabuchi *et al.*, Oncogene 8:1693-1696 [1993]; and Brent W0 94/10300; each of which is herein incorporated by reference), to identify other proteins, that bind to or interact with MCFD2 ("MCFD2-binding proteins" or "MCFD2-
5 bp") and are involved in MCFD2 activity. Such MCFD2-bps can be activators or inhibitors of signals by the MCFD2 proteins or targets as, for example, downstream elements of a MCFD2-mediated signaling pathway.

Modulators of MCFD2 expression can also be identified. For example, a cell or cell free mixture is contacted with a candidate compound and the expression of MCFD2
10 mRNA or protein evaluated relative to the level of expression of MCFD2 mRNA or protein in the absence of the candidate compound. When expression of MCFD2 mRNA or protein is greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of MCFD2 mRNA or protein
expression. Alternatively, when expression of MCFD2 mRNA or protein is less (i.e.,
15 statistically significantly less) in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of MCFD2 mRNA or protein expression. The level of MCFD2 mRNA or protein expression can be determined by methods described herein for detecting MCFD2 mRNA or protein.

A modulating agent can be identified using a cell-based or a cell free assay, and
20 the ability of the agent to modulate the activity of a MCFD2 protein can be confirmed *in vivo*, e.g., in an animal such as an animal model for a disease (e.g., an animal with hematologic disease; See e.g., Hildenbrandt and Otto, J. Am. Soc. Nephrol. 11:1753 [2000]).

25 C. Therapeutic Agents

This invention further pertains to novel agents identified by the above-described screening assays. Accordingly, it is within the scope of this invention to further use an agent identified as described herein (e.g., a MCFD2 modulating agent or mimetic, a MCFD2 specific antibody, or a MCFD2-binding partner) in an appropriate animal model
30 (such as those described herein) to determine the efficacy, toxicity, side effects, or mechanism of action, of treatment with such an agent. Furthermore, novel agents

identified by the above-described screening assays can be, *e.g.*, used for treatments of hematologic disease (*e.g.*, including, but not limited to, F5F8D).

IX. Pharmaceutical Compositions Containing MCFD2 Nucleic Acid, Peptides, and Analogs

The present invention further provides pharmaceutical compositions which may comprise all or portions of MCFD2 polynucleotide sequences, MCFD2 polypeptides, inhibitors or antagonists of MCFD2 bioactivity, including antibodies, alone or in combination with at least one other agent, such as a stabilizing compound, and may be administered in any sterile, biocompatible pharmaceutical carrier, including, but not limited to, saline, buffered saline, dextrose, and water.

The methods of the present invention find use in treating diseases or altering physiological states characterized by mutant MCFD2 alleles (*e.g.*, F5F8D). Peptides can be administered to the patient intravenously in a pharmaceutically acceptable carrier such as physiological saline. Standard methods for intracellular delivery of peptides can be used (*e.g.*, delivery via liposome). Such methods are well known to those of ordinary skill in the art. The formulations of this invention are useful for parenteral administration, such as intravenous, subcutaneous, intramuscular, and intraperitoneal. Therapeutic administration of a polypeptide intracellularly can also be accomplished using gene therapy as described above.

As is well known in the medical arts, dosages for any one patient depends upon many factors, including the patient's size, body surface area, age, the particular compound to be administered, sex, time and route of administration, general health, and interaction with other drugs being concurrently administered.

Accordingly, in some embodiments of the present invention, MCFD2 nucleotide and MCFD2 amino acid sequences can be administered to a patient alone, or in combination with other nucleotide sequences, drugs or hormones or in pharmaceutical compositions where it is mixed with excipient(s) or other pharmaceutically acceptable carriers. In one embodiment of the present invention, the pharmaceutically acceptable carrier is pharmaceutically inert. In another embodiment of the present invention,

MCFD2 polynucleotide sequences or MCFD2 amino acid sequences may be administered alone to individuals subject to or suffering from a disease.

Depending on the condition being treated, these pharmaceutical compositions may be formulated and administered systemically or locally. Techniques for formulation and administration may be found in the latest edition of "Remington's Pharmaceutical Sciences" (Mack Publishing Co, Easton Pa.). Suitable routes may, for example, include oral or transmucosal administration; as well as parenteral delivery, including intramuscular, subcutaneous, intramedullary, intrathecal, intraventricular, intravenous, intraperitoneal, or intranasal administration.

For injection, the pharmaceutical compositions of the invention may be formulated in aqueous solutions, preferably in physiologically compatible buffers such as Hanks' solution, Ringer's solution, or physiologically buffered saline. For tissue or cellular administration, penetrants appropriate to the particular barrier to be permeated are used in the formulation. Such penetrants are generally known in the art.

In other embodiments, the pharmaceutical compositions of the present invention can be formulated using pharmaceutically acceptable carriers well known in the art in dosages suitable for oral administration. Such carriers enable the pharmaceutical compositions to be formulated as tablets, pills, capsules, liquids, gels, syrups, slurries, suspensions and the like, for oral or nasal ingestion by a patient to be treated.

Pharmaceutical compositions suitable for use in the present invention include compositions wherein the active ingredients are contained in an effective amount to achieve the intended purpose. For example, an effective amount of MCFD2 may be that amount that suppresses coagulopathy. Determination of effective amounts is well within the capability of those skilled in the art, especially in light of the disclosure provided herein.

In addition to the active ingredients these pharmaceutical compositions may contain suitable pharmaceutically acceptable carriers comprising excipients and auxiliaries that facilitate processing of the active compounds into preparations that can be used pharmaceutically. The preparations formulated for oral administration may be in the form of tablets, dragees, capsules, or solutions.

The pharmaceutical compositions of the present invention may be manufactured in a manner that is itself known (*e.g.*, by means of conventional mixing, dissolving, granulating, dragee-making, levigating, emulsifying, encapsulating, entrapping or lyophilizing processes).

5 Pharmaceutical formulations for parenteral administration include aqueous solutions of the active compounds in water-soluble form. Additionally, suspensions of the active compounds may be prepared as appropriate oily injection suspensions. Suitable lipophilic solvents or vehicles include fatty oils such as sesame oil, or synthetic fatty acid esters, such as ethyl oleate or triglycerides, or liposomes. Aqueous injection
10 suspensions may contain substances that increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Optionally, the suspension may also contain suitable stabilizers or agents that increase the solubility of the compounds to allow for the preparation of highly concentrated solutions.

 Pharmaceutical preparations for oral use can be obtained by combining the active
15 compounds with solid excipient, optionally grinding a resulting mixture, and processing the mixture of granules, after adding suitable auxiliaries, if desired, to obtain tablets or dragee cores. Suitable excipients are carbohydrate or protein fillers such as sugars, including lactose, sucrose, mannitol, or sorbitol; starch from corn, wheat, rice, potato, etc; cellulose such as methyl cellulose, hydroxypropylmethyl-cellulose, or sodium
20 carboxymethylcellulose; and gums including arabic and tragacanth; and proteins such as gelatin and collagen. If desired, disintegrating or solubilizing agents may be added, such as the cross-linked polyvinyl pyrrolidone, agar, alginic acid or a salt thereof such as sodium alginate.

 Dragee cores are provided with suitable coatings such as concentrated sugar
25 solutions, which may also contain gum arabic, talc, polyvinylpyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures. Dyestuffs or pigments may be added to the tablets or dragee coatings for product identification or to characterize the quantity of active compound, (*i.e.*, dosage).

30 Pharmaceutical preparations that can be used orally include push-fit capsules made of gelatin, as well as soft, sealed capsules made of gelatin and a coating such as

glycerol or sorbitol. The push-fit capsules can contain the active ingredients mixed with a filler or binders such as lactose or starches, lubricants such as talc or magnesium stearate, and, optionally, stabilizers. In soft capsules, the active compounds may be dissolved or suspended in suitable liquids, such as fatty oils, liquid paraffin, or liquid polyethylene glycol with or without stabilizers.

Compositions comprising a compound of the invention formulated in a pharmaceutical acceptable carrier may be prepared, placed in an appropriate container, and labeled for treatment of an indicated condition. For polynucleotide or amino acid sequences of MCFD2, conditions indicated on the label may include treatment of condition related to coagulopathy or thrombosis.

The pharmaceutical composition may be provided as a salt and can be formed with many acids, including but not limited to hydrochloric, sulfuric, acetic, lactic, tartaric, malic, succinic, *etc.* Salts tend to be more soluble in aqueous or other protonic solvents that are the corresponding free base forms. In other cases, the preferred preparation may be a lyophilized powder in 1 mM-50 mM histidine, 0.1%-2% sucrose, 2%-7% mannitol at a pH range of 4.5 to 5.5 that is combined with buffer prior to use.

For any compound used in the method of the invention, the therapeutically effective dose can be estimated initially from cell culture assays. Then, preferably, dosage can be formulated in animal models (particularly murine models) to achieve a desirable circulating concentration range that adjusts MCFD2 levels.

A therapeutically effective dose refers to that amount of MCFD2 that ameliorates symptoms of the disease state. Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, *e.g.*, for determining the LD₅₀ (the dose lethal to 50% of the population) and the ED₅₀ (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index, and it can be expressed as the ratio LD₅₀/ED₅₀. Compounds that exhibit large therapeutic indices are preferred. The data obtained from these cell culture assays and additional animal studies can be used in formulating a range of dosage for human use. The dosage of such compounds lies preferably within a range of circulating concentrations that include the ED₅₀ with

little or no toxicity. The dosage varies within this range depending upon the dosage form employed, sensitivity of the patient, and the route of administration.

The exact dosage is chosen by the individual physician in view of the patient to be treated. Dosage and administration are adjusted to provide sufficient levels of the active moiety or to maintain the desired effect. Additional factors which may be taken into account include the severity of the disease state; age, weight, and gender of the patient; diet, time and frequency of administration, drug combination(s), reaction sensitivities, and tolerance/response to therapy. Long acting pharmaceutical compositions might be administered every 3 to 4 days, every week, or once every two weeks depending on half-life and clearance rate of the particular formulation.

Normal dosage amounts may vary from 0.1 to 100,000 micrograms, up to a total dose of about 1 g, depending upon the route of administration. Guidance as to particular dosages and methods of delivery is provided in the literature (*See*, U.S. Pat. Nos. 4,657,760; 5,206,344; or 5,225,212, all of which are herein incorporated by reference). Those skilled in the art will employ different formulations for MCFD2 than for the inhibitors of MCFD2. Administration to the bone marrow may necessitate delivery in a manner different from intravenous injections.

EXPERIMENTAL

The following examples are provided in order to demonstrate and further illustrate certain preferred embodiments and aspects of the present invention and are not to be construed as limiting the scope thereof.

METHODOLOGY

Subjects

Blood samples were obtained after informed consent from all individuals (study protocol approved by the institutional review board). Ten families, including 19 affected individuals in whom no LMAN1 mutations were detected, were used for genotyping (Figure 1). Families 1, 2, 3, 7 and 8 correspond to previously reported families A5, A7, A13, A21 and A28, respectively (Neerman-Arbez et al., 1999). Families 4, 5, 9 and 10 correspond to previously reported families 13, 16, 14 and 19 respectively (Nichols et al.,

1999). Family 6 is a large Turkish family not previously reported. An Epstein-Barr virus immortalized lymphoblast line was derived from individual IV:4 of this family, and Western blot analysis indicates a normal level of LMAN1 in a lymphoblast extract (Figure 3). Additional patient materials not included in the genome scan but included in the subsequent mutational analysis are two EBV immortalized lymphoblast lines derived from the probands of a previously reported family A29 (family 11) (Neerman-Arbez et al., 1999) and the first reported F5F8D family (family 12) (Oeri et al., 1954). Both lymphoblasts demonstrate normal levels of LMAN1 ((Neerman-Arbez et al., 1999) and Figure 3). Nine of the affected patients available for study are the progeny of first-cousin marriages (families 2-6) and one is the offspring of a marriage of first cousin once removed(family 1). There is no known consanguinity in families 7-12.

Linkage Analysis

A genome-wide linkage scan was performed with 382 polymorphic microsatellite markers spaced an average of 10 cM (panels 1-27 of the ABI Prism Linkage Mapping Set-MD 10 (Applied Biosystems) as described (Levy et al., 2001). Fine mapping was performed by PCR with labeled synthetic primer pairs for additional markers and analysis of PCR products on 8% denaturing polyacrylamide gels as described (Nichols et al., 1997). All markers were scored independently by two observers, the scores compared, and discrepancies resolved by re-examination of the chromatograms/gels. Linkage analysis was performed using MAPMAKER/HOMOZ (Kruglyak et al., 1995) as previously described (Nichols et al., 1997).

RT-PCR, RACE and Northern Blot Analysis

Total cellular RNA was prepared from EBV-transformed lymphoblasts and HeLa cells. The 5' end of the mRNA was determined by rapid amplification of cDNA ends (RACE) with a FirstChoice RLM-RACE kit (Ambion) using primers AGCAGGCCACACAGGAAG and CTCTTGGTCGTGCACTGTGT (SEQ ID NO:3). Sequence of RTPCS products was determined by the University of Michigan DNA Sequencing Core. A PCR product containing the coding sequence of the *MCFD2* cDNA

was amplified from a human MOLT4 T-cell cDNA library as previously described (Levy et al., 2001) using primers GCTTGGTACCTGCAGTGATTTTGCAAATTCAG and GGACTCGAGACCATGAGATCCCTGCTCAGA (SEQ ID NO 4). After gel purification, this DNA fragment was labeled by random priming and used to hybridize to a FirstChoice Human Northern Blot (Ambion) in Rapid-Hyb buffer (Amersham) according to the manufacturer's specifications. The same pair of primers was used to screen *MCFD2* expression by PCR on cDNA obtained from a multiple tissue cDNA panel (Clontech).

Gene-specific probes were used to detect the mRNA of individual candidate genes on Northern blots of 20 mg total RNA extracted from EBV-immortalized cell lines of two normal and four affected individuals.

Construction of Expression Vectors

To construct an *E. coli* thioredoxin fusion expression vector, the coding sequence of the wild-type *MCFD2* gene was amplified from a human T-cell cDNA library as above using primers GGACTCGAGACCATGAGATCCCTGCTCAGA (SEQ ID NO:20) and GGTGGAATTCTGCAGTGATTTTGCAAATTCAG (SEQ ID NO:21), digested with *EcoR1* and *Xho1* and subcloned into pThioHis A (Invitrogen). To construct a *his*-tagged recombinant protein expression vector, the above plasmid was linearized with *EcoR1*, end filled with Klenow enzyme, and digested with *Xho1*. The resulting DNA fragment was subcloned into pET15b (Novagen) between the *Xho1* and *BamH1* (blunted) sites. For expression in mammalian cells, wild-type cDNA was amplified from the above human cDNA library using primers GGACTCGAGATATTGATGACCATGAGATCC (SEQ ID NO:22) and GCTTGGTACCTGCAGTGATTTTGCAAATTCAG (SEQ ID NO:23), digested with *Xho1* and *Asp718* and subcloned into pcDNA3.1/*myc-his* (Invitrogen). Mutant cDNA expression vectors were constructed by amplification from the linearized pcDNA3.1/*MCFD2-myc-his* using the same 5' primer as above and a 3' primer of either D129E (GCTTGGTACCTGCAGTGATTTTGCAAATTCAGCATAGTCAATGTATCCATCAT TGTTCTTCTCATCATCTCTCAAAACACCATC) (SEQ ID NO:24), I136T (GCTTGGTACCTGCAGTGATTTTGCAAATTCAGCATAGTCAGTGTATCCATCAT

TGTTCTTGTC) (SEQ ID NO:25) or I136V
(GCTTGGTACCTGCAGTGATTTTGCAAATTCAGCATAGTCAACGTATCCATCA
TTGTTCTTGTC) (SEQ ID NO:26), digested with *Xho*I and *Asp*718, and cloned into
pcDNA3.1/*myc-his*. Plasmid constructs were verified by DNA sequencing.

5

Antibodies

Antibodies were generated against full length *his*-tagged MCFD2 purified from
induced *E. coli* extracts by a His-Trap column (Amersham). The recombinant protein
was separated by SDS PAGE and used to generate rabbit antisera by a commercial
10 supplier (Invitrogen). MCFD2 antiserum was affinity purified using a purified *his*-tagged
MCFD2 coupled Amino-Link column (Pierce). Purified IgG was coupled to CNBr
activated Sepharose 4B beads (Amersham) according to manufacturer's instructions. *His*
tagged MCFD2 was also used to generate monoclonal antibodies at the University of
Michigan Hybridoma Core Facility. Hybridoma clones were screened by direct ELISA
15 using plates coated with purified *his*-tagged MCFD2, and confirmed by Western blot
analysis.

Immunofluorescence

Immunofluorescence of COS-1 cells transfected with *MCFD2* expression vectors
20 was performed as described previously (Nichols et al., 1998), except that
permeabilization was achieved by incubating the slides in 0.5% Triton X-100 for 5 min.
Images were visualized and captured using an Olympus Fluoview confocal microscope.
Primary antibodies were a monoclonal anti-*myc* antibody (*sc*-789, Santa Cruz
Biotechnology), a monoclonal anti-LMAN1 antibody (G1/93, a gift from H.P. Hauri,
25 University of Basel, Switzerland), and a monoclonal anti-PDI antibody (ABR).
Secondary antibodies were FITC-conjugated goat-anti-mouse IgG and Texas red
conjugated goat-anti-rabbit IgG (Molecular Probes).

Metabolic Labeling, Immunoprecipitation and Protein Analysis

30 COS-1 cells were transfected using the DEAE/Dextran method (Sussman and
Milman, 1984) and metabolically labeled at 48 hours post transfection for 30 min with

[35S] methionine (250 mCi/ml in methionine-free medium). Immunoprecipitations were performed using equal TCA precipitable counts of labeled cell extracts as described previously (Moussalli et al., 1999). Proteins were analyzed by SDS-PAGE under reducing conditions and visualized by fluorography. Immunoprecipitations were also performed in extracts of unlabeled cells and analyzed by Western blot analysis.

EXAMPLE 1

A second locus for combined deficiency of factor V and factor VIII maps to human chromosome 2p21-2p16.3

Earlier analysis of multiple F5F8D patients identified a subset of approximately 30% in whom no *LMAN1* mutation could be detected, with strong evidence suggesting a second locus (Neerman-Arbez et al., 1999; Nichols et al., 1999).

In the absence of strong candidates for the second F5F8D gene, whole genome linkage analysis was performed by homozygosity mapping (Lander and Botstein, 1987) similar to the strategy that was used to identify *LMAN1* (Nichols et al., 1997; Nichols et al., 1998). Ten pedigrees available to us for genetic study are shown in Figure 1. DNA prepared from the 19 available affected individuals was genotyped for a total of 382 polymorphic markers spaced approximately 10 cM apart, spanning all 22 autosomes. The most significant increase in homozygosity was observed for the markers D2S391 and D2S337 on the long arm of chromosome 2 (maximum LOD of 3.59 for the interval between these two markers by MAPMAKER/HOMOZ (Kruglyak et al., 1995)).

The probands and all available family members were then genotyped using a large panel of polymorphic markers known to map to this region (research.marshfieldclinic.org/genetics/, cephb.fr/ceph-genethon-map, lpg.nci.nih.gov/html-chlc/ChlcMarkers web sites). The results allowed the gene to be placed in the ~2.6 cM interval between markers D2S391 and D2S2739. Additional markers were developed from the draft human genome sequence (Lander et al., 2001) (<http://genome.ucsc.edu>) as previously described (Levy et al., 2001). This analysis mapped the disease locus to a 1.8 cM minimal non-recombinant interval between markers

BZ31 and BZ18 on chromosome 2p21-2p16.3 (Figure 2A). Multi-point linkage analysis demonstrates a maximum LOD score of 6.4 at D2S2227. The two affected siblings in family 8 inherited different allele combinations from their parents (Figure 1). In addition, although family 3 is consanguineous, the affected individual (IV1) is heterozygous throughout the region, with two of the four unaffected siblings (IV4 and IV5) homozygous for an allele shared with the affected sib (Figure 1 and additional markers, data not shown). These results suggest that the gene responsible for F5F8D is not located in this region of the genome in family 8, and probably also family 3. With these two families excluded, the maximum LOD score is 10.2 at marker D2S2227.

EXAMPLE 2

Mutations in *MCFD2* cause combined deficiency of factor V and factor VIII

Analysis of the draft human genome sequence demonstrates that the candidate genetic interval between BZ31 and BZ18 is about 2.4 Mb with four major gaps (Figure 2B). Known genes in the interval include endothelial PAS domain protein 1, phosphatidylinositol glycan class F, calmodulin 2, TC10a, *mutS* homologues 2 and 6, and T-cell leukemia virus enhancer factor (GenBank accession numbers NM_001430, NM_002643, NM_001743, NM_012249, NM_000251, NM_000179, and NM_002158, respectively). This region also contains ~12 predicted genes (Figure 2B) and a number of partial ESTs. None of these potential genes exhibit sequence homology to known resident proteins of the ER or the Golgi apparatus. Sixteen candidate genes/partial ESTs were initially screened by Northern blot analysis of patient total lymphoblast RNA. No significant alterations in mRNA size or abundance were observed for the nine predicted genes for which a signal was detected in lymphoblasts (data not shown, GenBank accession numbers for these genes are: NM_001743, NM_002643, NM_012249, XM_031626, NM_014171, AF073958, XM_065167, AF174599 and M23161).

Individual candidate genes were next screened by direct DNA sequencing. DNA fragments containing coding exons and exon-intron junctions were amplified by PCR

from patient DNAs or amplified by RT-PCR from patient lymphoblast RNA. A total of eight genes (*CALM2*, *PIG-F*, *TC-10*, *KIAA1140*, *HSPC139*, *CISH6*, *FBX11*, and *HUMTRANSC*. GenBank accession numbers: NM_001743, NM_002643, NM_012249, XM_031626, NM_014171, AF073958, AF174599, and M23161, respectively) were
5 either completely or partially sequenced. Candidate mutations were identified in the predicted exons and intron/exon junctions of *HUMTRANSC*. This partial cDNA was previously reported as a transposonlike element RNA, with an unusually long 3' untranslated region containing a transposonlike human repeat element, THE 1 (Deka et al., 1988). The *HUMTRANSC* sequence in 10 GenBank contains a number of errors. We
10 have deposited the correct full-length cDNA (~4.1 kb) sequence into GenBank (accession number AF537214). This gene, which we renamed *MCFD2*, spans 4 exons encompassing ~19 kb in the human genome and contains a 145 amino acid open reading frame, predicting a 16 kDa protein. Notable features of the amino acid sequence include a predicted signal peptide at the N-terminus and two calmodulin-like EF-hands for
15 putative calcium binding at the C-terminus (Figure 2C).

The full coding sequence and all intron/exon junctions of *MCFD2* were amplified by PCR from genomic DNA obtained from twelve F5F8D families. Seven distinct *MCFD2* mutations were identified, accounting for 9 of the 12 families (Table 1). Three of the 7 identified mutations result in frameshifts, including single nucleotide deletions in
20 exon 2 (103delC) and exon 3 (249delT), and an 8 nucleotide deletion in exon 3 (263-270delTTGATGGC). Splice site mutations were identified in 4 families: a G to A substitution (309+1G->A) in the invariant GT of the intron 3 splice donor site (families 1 and 5), and a G to A substitution (149+5G->A) in the extended donor splice site consensus sequence of intron 2 (families 10 and 12). The remaining two mutations,
25 found in families 4 and 9 respectively, result in single amino acid substitutions in the second putative EF-hand domain (D129E and I136T). These two missense mutations were excluded as common sequence polymorphisms by screening a panel of over 200 unaffected chromosomes. No *MCFD2* mutations were identified in families 2, 3, and 8. These data are consistent with the evidence noted above that F5F8D in families 3 and 8 is
30 not linked to *MCFD2*.

Table 1
MCFD2 Mutations in Families with Combined Deficiency of Factor V and Factor VIII

Nucleotide	Amino Acid	Family	Nucleic Acid SEQ ID NO	Amino Acid SEQ ID NO
149 +5G> A	Donor splice site	10, 12	5	6
309 +1G> A	Donor splice site	1, 5	7	8
103delC	Frameshift at aa 35	7	9	10
249delT	Frameshift at aa 83	6	11	12
263-270delTTGATGGC	Frameshift at aa 88	11	13	14
C387> G	D129E	4	15	16
T407> C	I136T	9	17	18

5

EXAMPLE 3

MCFD2 mutations result in loss of protein expression

10 Though cells from families 1 and 5 were not available, the 309+1G->A mutation abolishes the invariant splice donor consensus and would be expected to result in a null allele. To test the effect of the 242+5G->A mutation on mRNA splicing, RT-PCR was performed on total RNA prepared from patient and control lymphoblasts. This mutation appears to eliminate utilization of the normal intron 2 splice donor with skipping of exon

15 2, which contains the start codon. In addition, a minor product was observed from the activation of a cryptic donor splice site at +161 of intron 2 that results in the insertion of 17 amino acids followed by a stop codon (data not shown).

 To determine whether the mutations identified in the *MCFD2* gene result in loss of MCFD2 expression, Western blot analysis of extracts from available EBV-transformed

20 lymphoblast lines was performed, using an anti-human MCFD2 monoclonal antibody (Figure 3B). A 16-kDa band, consistent with the predicted molecular mass of MCFD2, was detected in lymphoblasts derived from normal individuals, but was absent in cells derived from four patients with *MCFD2* null mutations (Figure 3B).

25

EXAMPLE 4

***MCFD2* is Widely Expressed and Co-localizes with LMAN1**

5 A full-length, ~4.1 kb *MCFD2* mRNA was detected in multiple tissues by Northern blot analysis (Figure 4A), with *MCFD2* expression evident by RT-PCR in additional tissues. Variation in expression was noted, with lower levels seen in brain and lung (Figure 4A). In addition, potential 0.8 kb, 1.2 kb and 1.8 kb mRNA species were detected in various tissues that could represent alternate spliced forms. The same
10 *MCFD2* coding sequence associated with a shorter mRNA was recently deposited in GenBank as a novel secretory molecule with neuronal survival effect from adult rat hippocampal stem cells, designated as SDNSF (accession number NM_139279). Multiple potential splice forms can also be predicted from available SDNSF EST sequences (ncbi.nlm.nih.gov/locuslink web site). Three potential *MCFD2* transcriptional
15 start sites were identified by a rapid amplification of cDNA ends (RACE) procedure that specifically amplifies the 5' capped mRNA (Figure 4B). DNA upstream of the transcriptional start is rich in GC and does not contain a TATA box-like element, consistent with the promoter of a housekeeping-type gene.

To determine the subcellular localization of *MCFD2*, cDNA containing the
20 complete *MCFD2* coding sequence of the wild-type, D129E mutant, I136T mutant and the similar I136V substitution were cloned into pcDNA3.1/*myc-his*, yielding recombinant proteins with *myc* and *his* epitopes fused to the C-terminus of *MCFD2*. Transiently transfected HeLa cells were stained simultaneously with a rabbit polyclonal antiserum to the *c-myc* epitope, and either a monoclonal antibody to LMAN1, or a monoclonal
25 antibody to protein disulfide isomerase (PDI), an ER marker. A distinct pattern was observed with anti-*myc* antibody in cells expressing wild-type *MCFD2*, with nearly complete overlap with anti-LMAN1 staining and minimal overlap with anti-PDI staining (Figure 5), demonstrating the co-localization of *MCFD2* and LMAN1 within the ERGIC.

In contrast, the D129E and I136T mutants both displayed a diffuse staining
30 pattern, which partially overlaps both LMAN1 and PDI, suggesting a broader distribution throughout the cell. *MCFD2* carrying the more conservative I136V substitution exhibits

a pattern more closely resembling wild type, though with a slight increase in ER staining (Figure 5).

EXAMPLE 5

5

Intracellular Accumulation of MCFD2 Requires LMAN1

As previously reported (Nichols et al., 1998; Neerman-Arbez et al., 1999), LMAN1 was readily detected by Western blot analysis of lymphoblasts derived from normal controls and from patients with *MCFD2*-null mutations, but was absent from LMAN1-null lymphoblasts (Figure 3A). Similarly, MCFD2 was readily detected in wildtype lymphoblasts but was absent from lymphoblasts derived from patients with *MCFD2*-null mutations. Unexpectedly, MCFD2 was also absent from LMAN1 null lymphoblasts (Figure 3B). A trace amount of MCFD2 was however detected upon immunoprecipitation from a large number of LMAN1-null cells (Figure 3C), consistent with a lack of retention and/or degradation of MCFD2 in the absence of LMAN1.

EXAMPLE 6

20

MCFD2 Interacts with LMAN1 in a Calcium-Dependent Manner

Since mutations in *MCFD2* and *LMAN1* cause identical F5F8D phenotypes, we examined the potential for interaction between these gene products. COS-1 cells were transiently transfected with combinations of expression vectors encoding *myc*-tagged wild-type or mutant MCFD2 or LMAN1 and metabolically labeled with [35S]-methionine.

Cell extracts were prepared and immunoprecipitated with antibodies against *c-myc* and LMAN1. Immunoprecipitation of lysates from cells transfected with wild-type *myc*-MCFD2 using antibody to *c-myc* detected a species migrating at approximately 18 kDa (Figure 6A, lane 3) that was not present in lysates from cells that did not receive DNA (Figure 6A, lanes 1). The 18 kDa species is the expected molecular mass for *myc*-

tagged MCFD2. A 53 kDa species co-precipitated with *myc*-MCFD2 that co-migrated with endogenous and over-expressed LMAN1 (Figure 6A, lane 3, and Figure 6B, lane 3). Anti-LMAN1 also co-precipitated *myc*-MCFD2 from cells transfected with *myc*-MCFD2 (Figure 6A, lane 4, and Figure 6B, lane 4). In addition, a faint, slightly faster migrating species co-precipitated with LMAN1 from all cell lysates (Figure 6A, lanes 2, 4, 6, 8, 10), including mock-transfected cells. This 16 kDa species is the expected molecular mass for endogenous MCFD2. The same band was also detected by immunoprecipitation of these cell lysates with a monoclonal anti-MCFD2 antibody (data not shown). These findings demonstrate that endogenous LMAN1 interacts with both endogenous MCFD2 as well as over-expressed MCFD2.

In contrast to these observations with wild-type MCFD2, co-immunoprecipitation of *myc*-tagged MCFD2 and LMAN1 was not observed from lysates of cells transfected with MCFD2 carrying either the D129E or I136T mutations (Figure 6A, lanes 5, 6, 9, 10), although the proteins were synthesized at high levels. The more conservative I136V substitution reduced, but did not eliminate LMAN1 interaction under these experimental conditions (Figure 6A, lanes 7, 8).

MCFD2 contains EF-hand domains that may bind calcium. To study the role of calcium in MCFD2 function, lysates were prepared from cells co-transfected with LMAN1 and *myc*-MCFD2. Treatment of cell lysates with the calcium-specific chelator EGTA prior to immunoprecipitation destroyed the LMAN1 interaction with *myc*-MCFD2 (Figure 6B, lanes 5-6). Recalcification of the EGTA-treated lysates reconstituted the MCFD2 association with LMAN1 (Figure 6B, lanes 7-8). These results demonstrate that MCFD2 associates with LMAN1, and that this interaction is calcium-dependent.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are

obvious to those skilled in the relevant fields are intended to be within the scope of the following claims.

REFERENCES

5

Appenzeller,C., Andersson,H., Kappeler,F., and Hauri,H.P. (1999). The lectin ERGIC-53 is a cargo transport receptor for glycoproteins. *Nat. Cell Biol.* 1, 330-334.

10

Arar,C., Carpentier,V., Le Caer,J.P., Monsigny,M., Legrand,A., and Roche,A.C. (1995). ERGIC-53, a membrane protein of the endoplasmic reticulum-Golgi intermediate compartment, is identical to MR60, an intracellular mannose-specific lectin of myelomonocytic cells. *J. Biol. Chem.* 270, 3551-3553.

15

Barker,D., Schafer,M., and White,R. (1984). Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. *Cell* 36, 131-138.

Belden,W.J. and Barlowe,C. (2001). Role of Erv29p in collecting soluble secretory proteins into ER-derived transport vesicles. *Science* 294, 1528-1531.

20

Chiu,H.C., Schick,P.K., and Colman,R.W. (1985). Biosynthesis of factor V in isolated guinea pig megakaryocytes. *J. Clin. Invest.* 75, 339-346.

25

Deka,N., Wong,E., Matera,A.G., Kraft,R., Leinwand,L.A., and Schmid,C.W. (1988). Repetitive nucleotide sequence insertions into a novel calmodulin- related gene and its processed pseudogene. *Gene* 71, 123-134.

Do,H., Healey,J.F., Waller,E.K., and Lollar,P. (1999). Expression of factor VIII by murine liver sinusoidal endothelial cells. *J. Biol. Chem.* 274, 19587-19592.

30

Fiedler,K. and Simons,K. (1994). A putative novel class of animal lectins in the secretory pathway homologous to leguminous lectins. *Cell* 77, 625-626.

- Fiedler,K. and Simons,K. (1996). Characterization of VIP36, an animal lectin homologous to leguminous lectins. *J. Cell Sci. 109 (Pt 1)*, 271-276.
- 5 Ginsburg,D. (2002). Hemophilia and Other Disorders of Hemostasis. In Emery and Rimoin's Principles and Practice of Medical Genetics, Vol. II, D.L.Rimoin, J.M.Connor, and R.E.Pyeritz, eds. (New York: Churchill Livingstone), pp. 1926-1958.
- 10 Hebert,D.N., Foellmer,B., and Helenius,A. (1995). Glucose trimming and reglucosylation determine glycoprotein association with calnexin in the endoplasmic reticulum. *Cell 81*, 425-433.
- 15 Itin,C., Roche,A.C., Monsigny,M., and Hauri,H.P. (1996). ERGIC-53 is a functional mannose-selective and calcium-dependent human homologue of leguminous lectins. *Mol. Biol. Cell 7*, 483-493.
- 20 Kappeler,F., Klopfenstein,D.R., Foguet,M., Paccaud,J.P., and Hauri,H.P. (1997). The recycling of ERGIC-53 in the early secretory pathway. ERGIC-53 carries a cytosolic endoplasmic reticulum-exit determinant interacting with COPII. *J. Biol. Chem. 272*, 31801-31808.
- Kaufman,R.J. (1998). Post-translational modifications required for coagulation factor secretion and function. *Thromb. Haemost. 79*, 1068-1079.
- 25 Kaufman,R.J., Wasley,L.C., and Dorner,A.J. (1988). Synthesis, processing, and secretion of recombinant human factor VIII expressed in mammalian cells. *J. Biol. Chem. 263*, 6352-6362.
- 30 Klumperman,J., Schweizer,A., Clausen,H., Tang,B.L., Hong,W., Oorschot,V., and Hauri,H.P. (1998). The recycling pathway of protein ERGIC-53 and dynamics of the ERGolgi intermediate compartment. *J. Cell Sci. 111 (Pt 22)*, 3411-3425.

Kruglyak,L., Daly,M.J., and Lander,E.S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am. J. Hum. Genet.* 56, 519-527.

5

Kuehn,M.J., Herrmann,J.M., and Schekman,R. (1998). COPII-cargo interactions direct protein sorting into ER-derived transport vesicles. *Nature* 391, 187-190.

Lander,E.S. and Botstein,D. (1987). Homozygosity mapping: A way to map human
10 recessive traits with the DNA of inbred children. *Science* 236, 1567-1570.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K.,
Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., Heaford,A.,
Howland,J., Kann,L., Lehoczky,J., LeVine,R., McEwan,P., McKernan,K., Meldrim,J.,
15 Mesirov,J.P., Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M., Santos,R.,
Sheridan,A., Sougnez,C., Stange-Thomann,N., Stojanovic,N., Subramanian,A.,
Wyman,D., Rogers,J., Sulston,J., Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C.,
Carter,N., Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I., Durbin,R.,
French,L., Grafham,D., Gregory,S., Hubbard,T., Humphray,S., Hunt,A., Jones,M.,
20 Lloyd,C., McMurray,A., Matthews,L., Mercer,S., Milne,S., Mullikin,J.C., Mungall,A.,
Plumb,R., Ross,M., Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W.,
McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A., Chinwalla,A.T., Pepin,K.H.,
Gish,W.R., Chissole,S.L., Wendl,M.C., Delehaunty,K.D., Miner,T.L., Delehaunty,A.,
Kramer,J.B., Cook,L.L., Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T.,
25 Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T., Doggett,N., Cheng,J.F.,
Olsen,A., Lucas,S., Elkin,C., Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M.,
Scherer,S.E., Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H.,
Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L., Weinstock,G.M., Sakaki,Y.,
Fujiyama,A., Hattori,M., Yada,T., Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H.,
30 Totoki,Y., Taylor,T., Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P.,
Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R., Doucette-Stamm,L.,

- Rubenfield,M., Weinstock,K., Lee,H.M., Dubois,J., Rosenthal,A., Platzner,M.,
Nyakatura,G., Taudien,S., Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L.,
Rowen,L., Madan,A., Qin,S., Davis,R.W., Federspiel,N.A., Abola,A.P., Proctor,M.J.,
Myers,R.M., Schmutz,J., Dickson,M., Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R.,
5 Raymond,C., Shimizu,N., Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M.,
Schultz,R., Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R.,
McCombie,W.R., de la,B.M., Dedhia,N., Blocker,H., Hornischer,K., Nordsiek,G.,
Agarwala,R., Aravind,L., Bailey,J.A., Bateman,A., Batzoglu,S., Birney,E., Bork,P.,
Brown,D.G., Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M., Copley,R.R.,
10 Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S., Galagan,J., Gilbert,J.G., Harmon,C.,
Hayashizaki,Y., Haussler,D., Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S.,
Jones,T.A., Kasif,S., Kasprzyk,A., Kennedy,S., Kent,W.J., Kitts,P., Koonin,E.V., Korf,I.,
Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A., Mikkelsen,T., Moran,J.V., Mulder,N.,
Pollara,V.J., Ponting,C.P., Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E.,
15 Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L., Wallis,J., Wheeler,R.,
Williams,A., Wolf,Y.I., Wolfe,K.H., Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S.,
Peterson,J., Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., Szustakowski,J.,
de Jong,P., Catanese,J.J., Osoegawa,K., Shizuya,H., and Choi,S. (2001). Initial
sequencing and analysis of the human genome. *Nature* 409, 860-921.
- 20
- Levy,G.G., Nichols,W.C., Lian,E.C., Foroud,T., McClintick,J.N., McGee,B.M.,
Yang,A.Y., Siemieniak,D.R., Stark,K.R., Gruppo,R., Sarode,R., Shurin,S.B.,
Chandrasekaran,V., Stabler,S.P., Sabio,H., Bouhassira,E.E., Upshaw,J.D., Jr.,
Ginsburg,D., and Tsai,H.M. (2001). Mutations in a member of the ADAMTS gene family
25 cause thrombotic thrombocytopenic purpura. *Nature* 413, 488-494.
- Lewis,M.J., Sweet,D.J., and Pelham,H.R. (1990). The ERD2 gene determines the
specificity of the luminal ER protein retention system. *Cell* 61, 1359-1363.
- 30 Martin-Bermudo,M.D. and Brown,N.H. (2000). The localized assembly of extracellular
matrix integrin ligands requires cell-cell contact. *J. Cell Sci.* 113 Pt 21, 3715-3723.

- Martinez-Menarguez, J.A., Geuze, H.J., Slot, J.W., and Klumperman, J. (1999). Vesicular tubular clusters between the ER and Golgi mediate concentration of soluble secretory proteins by exclusion from COPI- coated vesicles. *Cell* 98, 81-90.
- 5 Moussalli, M., Pipe, S.W., Hauri, H.-P., Nichols, W.C., Ginsburg, D., and Kaufman, R.J. (1999). Mannose-dependent endoplasmic reticulum (ER)-Golgi intermediate compartment-53-mediated ER to Golgi trafficking of coagulation factors V and VIII. *J. Biol. Chem.* 274, 32539-32542.
- 10 Muniz, M., Morsomme, P., and Riezman, H. (2001). Protein sorting upon exit from the endoplasmic reticulum. *Cell* 104, 313-320.
- Muniz, M., Nuoffer, C., Hauri, H.P., and Riezman, H. (2000). The Emp24 complex recruits
15 a specific cargo molecule into endoplasmic reticulum-derived vesicles. *J. Cell Biol.* 148, 925-930.
- Munro, S. and Pelham, H.R. (1987). A C-terminal signal prevents secretion of luminal ER proteins. *Cell* 48, 899-907.
- 20 Neerman-Arbez, M., Johnson, K.M., Morris, M.A., McVey, J.H., Peyvandi, F., Nichols, W.C., Ginsburg, D., Rossier, C., Antonarakis, S.E., and Tuddenham, E.G.D. (1999). Molecular analysis of the ERGIC-53 gene in 35 families with combined factor V factor VIII deficiency. *Blood* 93, 2253-2260.
- 25 Nehls, S., Snapp, E.L., Cole, N.B., Zaal, K.J., Kenworthy, A.K., Roberts, T.H., Ellenberg, J., Presley, J.F., Siggia, E., and Lippincott-Schwartz, J. (2000). Dynamics and retention of misfolded proteins in native ER membranes. *Nat. Cell Biol.* 2, 288-295.
- 30 Nichols, W.C., Seligsohn, U., Zivelin, A., Terry, V.H., Arnold, N.D., Siemieniak, D.R., Kaufman, R.J., and Ginsburg, D. (1997). Linkage of combined factors V and VIII

deficiency to chromosome 18q by homozygosity mapping. *J. Clin. Invest.* 99, 596-601.

Nichols, W.C., Seligsohn, U., Zivelin, A., Terry, V.H., Hertel, C.E., Wheatley, M.A.,
Moussalli, M.J., Hauri, H.-P., Ciavarella, N., Kaufman, R.J., and Ginsburg, D. (1998).

- 5 Mutations in the ER-Golgi intermediate compartment protein ERGIC-53 cause combined
deficiency of coagulation factors V and VIII. *Cell* 93, 61-70.

Nichols, W.C., Terry, V.H., Wheatley, M.A., Yang, A., Zivelin, A., Ciavarella, N.,
Stefanile, C., Matsushita, T., Saito, H., de Bosch, N.B., Ruiz-Saez, A., Torres, A.,

- 10 Thompson, A.R., Feinstein, D.I., White, G.C., Negrier, C., Vinciguerra, C., Aktan, M.,
Kaufman, R.J., Ginsburg, D., and Seligsohn, U. (1999). ERGIC-53 gene structure and
mutation analysis in 19 combined factors V and VIII deficiency families. *Blood* 93,
2261-2266.

- 15 Nufer, O., Guldbrandsen, S., Degen, M., Kappeler, F., Paccaud, J.P., Tani, K., and
Hauri, H.P. (2002). Role of cytoplasmic C-terminal amino acids of membrane proteins in
ER export. *J. Cell Sci.* 115, 619-628.

- Oeri, J., Matter, M., Isenschmid, H., Hauser, F., and Koller, F. (1954). Angeborener mangel
20 an faktor V (parahaemophilie) verbunden mit echter haemophilie A bein zwei brudern.
Med. Probl. Paediatr. 1, 575-588.

- Pipe, S.W., Morris, J.A., Shah, J., and Kaufman, R.J. (1998). Differential interaction of
coagulation factor VIII and factor V with protein chaperones calnexin and calreticulin. *J.*
25 *Biol. Chem.* 273, 8537-8544.

- Prout, M., Damania, Z., Soong, J., Fristrom, D., and Fristrom, J.W. (1997). Autosomal
mutations affecting adhesion between wing surfaces in *Drosophila melanogaster*.
Genetics 146, 275-285.

30

Shekman, R. and Orci, L. (1996). Coat proteins and vesicle budding. *Science* 271, 1526-

1533.

Semenza, J.C., Hardwick, K.G., Dean, N., and Pelham, H.R. (1990). ERD2, a yeast gene required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway. *Cell* 61, 1349-1357.

Springer, S., Chen, E., Duden, R., Marzioch, M., Rowley, A., Hamamoto, S., Merchant, S., and Schekman, R. (2000). The p24 proteins are not essential for vesicular transport in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A* 97, 4034-4039.

Sussman, D.J. and Milman, G. (1984). Short-term, high-efficiency expression of transfected DNA. *Mol. Cell Biol.* 4, 1641-1643.

Tisdale, E.J., Plutner, H., Matteson, J., and Balch, W.E. (1997). p53/58 binds COPI and is required for selective transport through the early secretory pathway. *J. Cell Biol.* 137, 581-593.

Vollenweider, F., Kappeler, F., Itin, C., and Hauri, H.P. (1998). Mistargeting of the lectin ERGIC-53 to the endoplasmic reticulum of HeLa cells impairs the secretion of a lysosomal enzyme. *J. Cell Biol.* 142, 377-389.

Warren, G. and Mellman, I. (1999). Bulk flow redux? *Cell* 98, 125-127.

Wieland, F.T., Gleason, M.L., Serafini, T.A., and Rothman, J.E. (1987). The rate of bulk flow from the endoplasmic reticulum to the cell surface. *Cell* 50, 289-300.